

# **Data-Based Methods for Historical Grammar and Lexicon Extraction in a Diachronic Corpus**

**Magisterarbeit zur Erlangung des akademischen Grades Magister Artium im Fach  
Germanistische Linguistik**

HUMBOLDT-UNIVERSITÄT ZU BERLIN  
PHILOSOPHISCHE FAKULTÄT II  
INSTITUT FÜR DEUTSCHE SPRACHE UND LINGUISTIK

Eingereicht von Amir Zeldes, Matrikelnummer 512587,  
geboren am 23.09.1980 in Jerusalem, Israel

Wissenschaftliche Betreuung:  
Frau Professor Dr. Anke Lüdeling und Herr Professor Dr. Jonas Kuhn

Berlin, den 17.09.2007

## Table of Contents

<b>Abstract.....</b>	<b>2</b>
<b>1 Introduction.....</b>	<b>3</b>
1.1 Historical Linguistics and Corpus Linguistics.....	4
1.2 Uses and Limitations of Bible Corpora .....	6
1.3 Parallel Corpora and Automatic Extraction of Correspondences.....	8
<b>2 Setting Up the Corpus .....</b>	<b>11</b>
2.1 The Texts .....	11
2.2 Tokenization .....	15
2.3 Morphophonological Tagging with Polimorph .....	19
2.4 The Tag-Set.....	20
<b>3 Morphological Suffix Change .....</b>	<b>26</b>
3.1 Preliminary Remarks on Polish Declensions and their Origins.....	27
3.2 Nominal Suffix Changes in Minimal Pairs .....	28
3.2.1 Masculine Personal Plural.....	30
3.2.2 Masculine Inanimate Genitives: u# versus a# .....	33
3.2.3 Instrumental Plural Masculine and Neuter .....	34
3.2.4 Nominative Plural Masculine .....	36
3.2.5 Analogical Nominative-Accusative Plural Feminine i-Stems .....	37
3.3 Overview of Systematic Nominal Suffix Workloads .....	38
3.4 Summary and Evaluation.....	41
<b>4 Changes in Verbal Lexis and Word Formation .....</b>	<b>44</b>
4.1 Choosing Items .....	44
4.2 Identifying Parallels .....	46
4.3 Changes in Verbs and Verb Substitution Types .....	47
4.3.1 Retained Verbal Lemmas.....	49
4.3.2 Prefix Changes .....	51
4.3.3 Stem Replacement with Prefix Retention .....	55
4.3.4 Stem Alteration .....	58
4.3.5 Total Substitution.....	61
4.3.6 Non-Verbal Lemmas.....	62
4.4 Summary and Evaluation.....	65
<b>5 Syntactic and Grammatical Change .....</b>	<b>69</b>
5.1 Reduction of Token Sequences.....	69
5.2 Decline of the Active Past Participle .....	71
5.3 Possessive Adjectives versus Genitival Possession.....	72
5.4 Passive Participle Copulas .....	75
5.5 Summary and Evaluation.....	79
<b>6 Conclusion and Outlook.....</b>	<b>81</b>
<b>7 Bibliography .....</b>	<b>87</b>
<b>Appendix A – List of Abbreviations.....</b>	<b>93</b>
<b>Appendix B – the <i>agr</i> Function.....</b>	<b>95</b>

## **Abstract**

This work examines automatic and semi-automatic methods for the extraction, classification and quantification of historical grammar and lexis correspondences from a parallel diachronic corpus. Two digitized versions of the Polish Gospel of Matthew taken from the Gdansk Bible, originally printed in 1606, and Warsaw Bible, first published in 1975, are used as the database for this case study. Parallel distributions and cooccurrences of morphological, lexical and grammatical elements in an annotated electronic corpus created from these materials are presented and analyzed in light of traditional accounts of Polish historical grammar.

Diese Arbeit untersucht automatische und halbautomatische Methoden der Extrahierung, Klassifizierung und Quantifizierung von historischen Grammatik- und Lexikkorrespondenzen. Zwei digitalisierte Versionen des Matthäusevangeliums auf Polnisch, zum einen aus der Danziger Bibel (zum ersten Mal 1606 gedruckt) und zum anderen aus der Warschauer Bibel (erschienen 1975), bilden die Grunddaten für diese Fallstudie. Parallele Distributionen und Kookkurrenzen von morphologischen, lexikalischen sowie grammatischen Elementen in einem aus diesen Daten aufbereiteten annotierten Korpus werden vorgestellt und im Hinblick auf traditionelle, historische Grammatiken der polnischen Sprache analysiert.

## 1 Introduction

Historical linguistics, which is primarily founded on the analysis of documents from older languages, is arguably the linguistic discipline bearing the greatest affinity to corpora – older language stages are only observable in and through them. In this vein, Matti Rissanen (to appear) opens a forthcoming article on historical linguistics and corpus linguistics with the following words:

*“The introduction of corpora has had a revolutionary effect on language studies in the last few decades. This is particularly true of historical linguistics, which has to rely on written sources only; introspection and native-speaker competence cannot be relied on in the study of the language of previous centuries and millennia”.*

However this seemingly self-evident truth appears not yet to have reached its full expression in the computer era. Towering works of traditional historical grammar, all deeply founded in corpus evidence in the form of a philologist’s deep knowledge of manuscripts and texts, are still the rule in historical linguistics, and electronic corpus-based studies still the exception. The main question motivating the present work is “can historical grammar be derived directly from an electronic corpus?” In this I understand historical grammar to mean a systematic description of the relationship between two diachronically disparate synchronic language stages, which contains a mapping between them, similar to that found in traditional volumes of historical grammar. This mapping could account for very diverse diachronic changes, e.g. in a language’s inflectional morphology, word formation, grammatical categories, syntactic constructions and more. I will also be interested in the historical development of lexis, which is sometimes not, or often only partly, included in historical grammars, and especially in its intersection with word formation.

The focus of this work is on a digital corpus-based approach, which means that I will be concerned with automatic and semi-automatic data driven methods for investigating digitized historical and modern texts and the relationship between them,

which can be evaluated against what we already know from traditional historical grammars. Which results of these grammars can be replicated and which not? Can we find facts that are overlooked in such accounts, and if so, what kind of facts and why? I will examine these questions on the case of Middle and Modern Polish, through an attempt to extract historical developments from a parallel diachronic corpus, an object whose advantages and limitations will be discussed in depth further on. The data for this corpus will be extracted from a small part of the Polish New Testament, namely two translations of the Gospel of Matthew, originally dating from 1606 and 1975.

The remainder of this introduction will be devoted to methodological issues in the use of corpora in general, and parallel and biblical corpora in particular. Chapter 2 will describe the corpus explored in this work and its preparation in detail. The following chapters then explore different areas of historical grammar: the distributions of nominal inflections in chapter 3; verb formation and lexical change in chapter 4; and syntactic and grammatical change in chapter 5. Chapter 6 concludes the present work and offers some directions for future study.

## **1.1 Historical Linguistics and Corpus Linguistics**

Corpus linguistics, like most of modern linguistics, in fact has its roots in historical linguistics: some of the earliest linguistic endeavors were rooted in attempts to understand the relationship between older, usually holy texts and contemporary language (especially with the aim of describing and prescribing an idealized archaic liturgical language, e.g. Pāṇini's Sanskrit grammar). In a sense, electronic corpus-based linguistics only takes the descriptive study of such 'corpus-languages' to the next level, by offering an unbiased (insofar as any empiric science can be unbiased), data-driven tool complementing human observation and intuition. In an ideal case, we might want to use corpora not only to test our existing models of a language, but even to use the data to suggest new directions for research. Conversely, as already mentioned, research in historical linguistics is also limited to written corpus data: we simply have no sources except for texts. This means that the main, and nontrivial methodological presupposition of corpus linguistics is already made: that conclusions can be drawn from sample data

that apply to the state of affairs in the abstract language system. But what constitutes a sample, and what exactly is the abstract language being sampled from?

According to Biber (1993: 243), although sample size (i.e. corpus size) is often considered the most important factor in corpus design, representativeness is the more important criterion for the evaluation of corpora. Representativeness is understood to refer to the “extent to which a sample includes the full range of variability in a population” (ibid.). The situation in historical corpora is often severely at odds with this ideal: our documentation is a fragmentary, selective and in essence accidentally preserved cross-section of a language stage which is not representative of an entire language. In some cases only translated texts survive and they are often religiously motivated (see section 1.3 on both these issues). Just as the language of the Bible today is not representative of ‘general language’<sup>1</sup>, it is reasonable to assume that it was not representative of the vernacular in earlier times; however, we often have only limited or even no means to prove this, and even if heterogeneous texts survive, they hardly ever preserve anything likely to resemble a balanced sample containing e.g. colloquial language.

Nonetheless, as Labov (1994: 11) puts it: “Historical Linguistics can [...] be thought of as the art of making the best use of bad data”. Since we cannot alter the selection of texts that we have, we must redefine the population that we are sampling. By conceding that a corpus is not representative of the general language, we can limit our statements to a subset of that language, of which our corpus may serve as a sample. In this case, two versions of a part of the Polish Bible will be serving as a sample of the biblical sublanguage in Middle and Modern Polish, and as we shall see, they may be considered to contain a sufficient range of variability in this language type for many purposes. The discussion on the relevance of results from this sublanguage for the general language can therefore be deferred for now, though I will return to it later in the evaluation of some findings in view of general historical and comparative grammars, which are based on wider-ranging data. In the absence of previous, comparable corpus-

---

<sup>1</sup> I will use this concept somewhat loosely as an umbrella term encompassing what is accepted by speakers as the common ground of the different varieties, registers, genres etc. in their language.

based work, these grammars will play the main role in controlling results obtained from the data and evaluating their plausibility.

## **1.2 Uses and Limitations of Bible Corpora**

In choosing to use a parallel diachronic corpus, which should also ideally be already digitized and freely available, the choice is immediately directed towards the Bible. While not optimal for many purposes, Bible corpora have been widely used in historical linguistics long before the advent of computer technology, not only because of the text's theological and cultural significance, but simply because the Bible (and in particular the Gospels) is one of the earliest sizable coherent texts documented for many (especially European) languages. The Bible also has a number of advantages for digital corpus studies (Resnik et al., 1999): the digital text is freely available in an unparalleled variety of languages, and has been repeatedly updated in various periods, making it ideal for comparative and diachronic studies (see also Cysouw and Wälchli, to appear). The dependable consistency of verse alignment between corpora, which will be put to use later in this study, is both effortless and more accurate than many automatic alignments – misalignment occurs in only a handful of cases (Resnik et al., 1999: 135) compared to average success rates between 90-95% for automated alignments (admittedly on sentence alignment tasks, more fine-grained than verse alignment, see Simard et al., 2000: 54-55). The care taken in translating the Bible also makes omissions relatively unlikely.

There are however many problematic issues in using Bible corpora for linguistics. The main objections are probably (cf. Resnik et al., 1999):

1. that it is a translated text especially prone to loan translations and foreign constructions which preserve the language of the source text. The source text is often itself a translation, meaning we have to reckon with several layers of this phenomenon;
2. that it is a semantically very marked text, whose special religious content bears only a limited similarity to the 'general language';
3. that biblical language is by its nature conservative, and therefore unsuitable for historical study.

The first two points are not independent of each other: many expressions that can be traced back to loan translations form part of the style of biblical language. As a consequence, once a loan construction has been accepted into the language through the text, it often becomes part of that language's native inventory, a fact which speakers are usually unaware of. Are the expressions *God fearing* or *to fear God* valid English phrases, or the use of the German word *halsstarrig* lit. 'stiff-necked' to mean 'obstinate, stubborn'? These all represent loan translations from Latin, but originally reaching as far back as Biblical Hebrew, where *יָרָא אֶת-יְהוָה* 'to fear God' had the sense 'to be devout', and *קָשָׁה-עֲרָף* 'hard-naped' meant 'refusing to bow' and hence 'stubborn' (cf. Eng. *a stiffnecked people* in the King James Bible). This is of course no different for Polish, which often uses entire phrases directly quoted from the Bible even in day-to-day contexts (see Koziara, 2001 for an analysis of Polish Biblical phraseology and its sources in different Bible translations). While modern biblical languages owe their existence at least in part to a sort of 'translationese' (i.e. language with peculiar properties stemming from its translated origin, see e.g. Baroni and Bernardini, 2006), the naturalization of many of these forms is hard to ignore.

Additionally, although the Bible (and in fact any text) has some idiosyncratic properties, it still shows considerable overlap with the so-called general language. Resnik et al. (1999: 147) compared the vocabulary of the Modern English New International Version of the Bible with the control vocabulary list used to write definitions in the *Longman Dictionary of Contemporary English* (Proctor, 1978), which is meant to represent the core vocabulary of the language most suitable for learners. The Bible corpus contained around 80% of the lemmas on the 2,200 word list, thus showing that the Bible's vocabulary did in fact cover central areas of modern language.

Nonetheless, as already discussed in section 1.1, the nature of the data to be used in this study requires that statements be limited, at least in the first instance, to a biblical sublanguage. Recognizing this sublanguage as a valid variety of Polish is justified insofar as it is recognized by Polish speakers as belonging to their language and interacts with standard language as well. In all likelihood, this situation also applies to other "biblical languages", as scriptural language has generally been a very influential part of many



languages, shaping their standard literary varieties (cf. the influence of Luther’s Bible on the development of standard German (Wolf, 1996)).

That said, it remains important to avoid what has been termed the “God’s truth fallacy” (Rissanen, 1989), which essentially means reliance on a corpus as representative of an entire language or even sublanguage, while disregarding its limitation in belonging to a certain time, place, genre and author. A single Bible translation may be only one in a set of existing translations in a language, and in fact, in a much larger set of possible translations that were never made, and which might have differed in many factors, including simply translation style. On the other hand, in a text with such normative influence as the Bible, where the choice of each item in the corpus makes it the *de facto* normative bearer of the Biblical meaning invested in a particular passage, the case is all the stronger for regarding a single text as a sublanguage in itself. Some of these difficulties can only be addressed by evaluating results against other materials and studies, though obviously a comparison of many texts (in this case various Bible translations) could make a valuable contribution (see chapter 6 for further discussion).

As for the third objection above regarding the text’s conservatism, it has been partly addressed already, in that any possible conservatism in biblical language is immediately part of the characteristics of the sublanguage under investigation. Furthermore, conservatism has some advantages for historical research of the ‘general language’: if a new version of a conservative text was forced to alter some element or construction, it is all the more likely that it was really no longer tolerated or comprehensible in contemporary language. Those elements that were changed may thus indicate central points in historical grammar or lexicography.

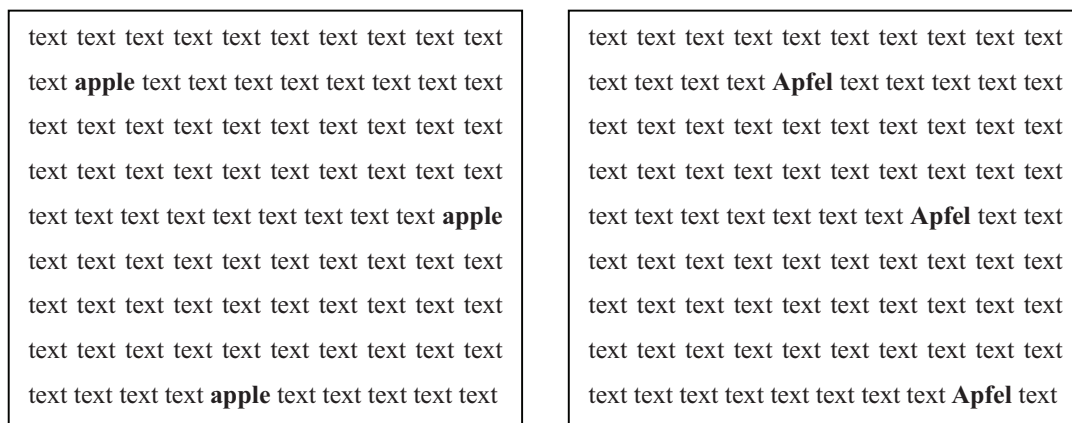
### **1.3 Parallel Corpora and Automatic Extraction of Correspondences**

While choosing to work with a parallel corpus limits corpus size and choice, there are considerable benefits as well. Firstly, the parallel content ensures that factors related to subject matter such as genre, register, domain etc. are completely comparable across the texts, making the diachronic disparity between them the central cause for any differences one might detect. The parallelism further extends to expected frequencies of grammatical functions, since, for example, the amount of participants in each predication in the

narrative is essentially constant. Thus if the text requires two human males to carry out a speech act directed at another one, the grammatical relations involved (conjoined, typically nominative subject, a matching predication, an experiencer, typically in dative case and an object clause containing direct speech) must be expressed in both texts, leading to a very high correspondence of lexical and morphological devices, which can easily reveal any subset of these elements which has changed. In chapter 3, this type of parallelism in distributions will be taken advantage of for the study of inflectional morphology.

Another type of parallelism that can be taken advantage of involves using the alignment between the corpora. As already mentioned, this can be done by using the Bible's given verse alignment. This alignment can be used to automatically calculate correlations between items in the different corpora, with the help of measures used in the automatic construction of parallel terminologies and in machine translation. While the application of these fields to historical linguistics may seem odd at first, it is not altogether unnatural – much like historical linguistics, they too are concerned with correspondences between two languages, giving answers to questions of the sort: “which Y in language B does X in language A correspond to? Under what circumstances, and how often?”.

The basic idea behind finding correspondences using alignment is that items that frequently appear in parallel segments are more likely to be translations of each other. This can be illustrated with the two text blocks in Figure 1.1, representing a parallel



**Fig. 1.1: Schematic representation of an English-German parallel text.**

English-German text. If we know that the lines in these two texts are aligned (meaning in our case that they belong to the same verse in each text), we can deduce that English *apple* is probably a translation of German *Apfel*, since the appearance of these words is correlated: the one appears in the one text if and only if the other appears in the other on the parallel line. While parallelism in natural texts is often less ideal than this one to one correspondence, the basic idea remains to find the most likely parallel based on such cooccurrences. Statistical techniques to deduce correspondences and their subsequent processing will be discussed in more depth in chapter 4.

While using cooccurrence measures for translating text automatically from older language stages into newer ones is of little interest in itself, it is possible to learn from the correspondences found in this way. A collection of correspondences between lexical items and expressions across language stages may be regarded as a historical dictionary, and correspondences between morphological features, grammatical categories and syntactic constructions that are annotated in the corpus can be seen as a sort of historical grammar. The advantage of automated techniques to find such correspondences is that hundreds of pairs of lexical items and constructions can be easily located and quantitatively evaluated, which would be difficult to do manually.

But caution is also required: things are rarely as simple as the pair *apple* : *Apfel* in historical linguistics. Finding some phenomenon in an old text and a different one under similar circumstances in a new text does not necessarily mean that one element has ‘replaced’ the other in the usual sense. Often two or more constructions or words compete for extended periods, having subtly different meanings and usage (cf. Rissanen, 1994: 27). Language change can be seen as a process characterized by variation or variability in the meaning and usage of different, yet related, competing linguistic signs (sometimes said to belong to a common ‘variant field’, cf. Curzan, 2003: 10). It can be expected that older elements will coexist with newer ones, and only gradually lose ground in certain contexts and senses. Only in this sense does a new attested form replace an old one: in being used in a corpus, one competitor is chosen over others within an overlapping field of possibilities, effectively taking part in the constant renegotiation of the linguistic value of the field and the items in it. I will revisit this problem in several places, especially in the evaluation of some of the results in chapter 4.

## 2 Setting Up the Corpus

### 2.1 The Texts

The corpus used in this study was created from digital versions of two Polish Bible translations of the Gospel of Matthew, and is called Polimatth. The choice of this particular Bible text was motivated by several factors. Firstly, the New Testament is a more suitable text with regard to the objections in section 1.2 than the Old Testament since it was written in relatively simple vernacular language, in order to facilitate its oral transmission to Christians and potential converts who were often illiterate (cf. Ehrman, 2005: 41-42), as opposed to the Old Testament text, which is a chronicle composed in a priestly, and not a popular context. The New Testament can therefore be expected to be more colloquial, and as such to share more developments with other forms of the language. The synoptic Gospels are particularly conducive to colloquial style and a variety of text types, as opposed to, say, epistles, which probably offer less varied text types. Especially valuable is the availability of both narrative text, usually restricted to third person past tense sentences, and direct speech, which admits the other tenses and persons, together with accompanying lexis (e.g. pronouns) and grammar, which may also bear the greatest resemblance to contemporary spoken language (cf. Taube, 1980: 121-122 for a similar reasoning). Finally, the Gospel of Matthew is the longest of all the Gospels, offering a larger corpus than the others, and approximately filling the limit of practicable length still allowing a good human proofreading of the annotation within the scope possible for the present work (on the need for proofreading see below).

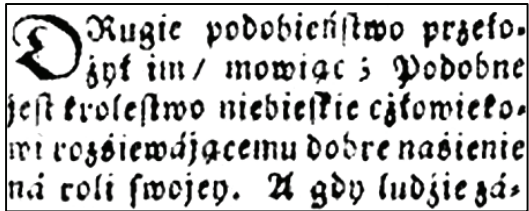
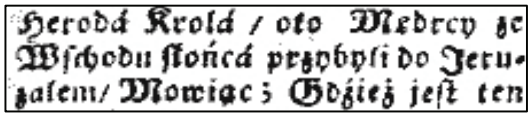
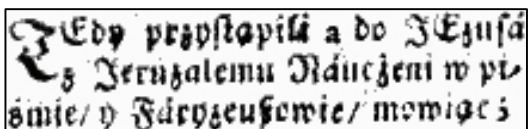

The older of the two translations was taken from the Protestant Gdansk Bible (Biblia Gdańska), first printed in 1606 (the New Testament) and then in 1632 (New and Old Testament), thus belonging in the Middle Polish period (ca. 16<sup>th</sup>-18<sup>th</sup> century). The choice of this particular text was motivated by several factors. Firstly, it is at once a relatively old yet complete Polish Bible translation, and since it was already printed, it has an authoritative established text (as opposed to conflicting manuscripts). Secondly, it served as the canonical Protestant Polish Bible into the 20<sup>th</sup> century, for which reason it may be considered a text that was at once an influential part of the Polish language, and

accepted as comprehensible Polish for a long time. Finally, although some Catholic Bibles are slightly older (e.g. the highly regarded late 16<sup>th</sup> century Bible of Jakub Wujek), there is no electronic version of them available on-line (except in scanned format for the Wujek Bible), which means a digitization effort would have been necessary.

However, for the Gdansk Bible too, the precise original text is not available electronically (except again in a scanned facsimile of an edition from 1632, available from the Württembergische Landesbibliothek Stuttgart at <http://www.bibliagdanska.pl>). A concession therefore had to be made to use the modern edition of the text, which is available on-line (originally obtained from <http://www.biblia.com.pl>, now available on the Polish Wikisource at [http://pl.wikisource.org/wiki/Biblia\\_Gda%C5%84ska](http://pl.wikisource.org/wiki/Biblia_Gda%C5%84ska)). This edition has undergone two revisions: in 1738 (the so-called Biblia Królewiecka) and in 1881 (the “Warsaw revision”). These revisions mostly affected the text’s orthography, punctuation, and in a few cases some inflectional endings (e.g. whether Hebrew proper names inflect or remain indeclinable)<sup>2</sup>. Importantly, the last revision has brought the text close enough in line with modern orthography to allow tagging the text using an adapted tagger of Modern Polish (more about which in section 2.3); the orthographic revision may therefore be regarded to some extent as a normalization of the underlying orthography. Figure 2.1a illustrates the overall faithfulness of the edition, while Figures 2.1b and 2.1c draw attention to orthographic change and regularization of suffixes for foreign names. Such differences are generally relatively minor, and infrequent enough to still allow a variety of linguistic studies of the text. The only frequent and systematic exception to this statement is the orthographic reform of neuter locative/instrumental singular and instrumental plural adjective suffixes, which were spelled differently from the masculine adjectives (despite widespread identical pronunciation) in the 19<sup>th</sup> century (Figure 2.1d). These adjectives take the endings <(i)em> and <(i)emi> instead of <(i/y)m> and <(i/y)mi> (see Klemensiewicz et al., 1955: 331-332, 336 for details). Though an exact transcription of the original text would have been of course theoretically desirable, it would have made tagging the text much more difficult, and in any case, the digitization effort was deemed impracticable for the scale of this work.

---

<sup>2</sup> Changes to the actual text are few, but also occur occasionally. Such special deviations will be identified using the aforementioned facsimile and noted in examples where they appear.

	<p>Drugie podobieństwo przełożył im, mówiąc: Podobne jest królestwo niebieskie człowiekowi, rozsiewającemu dobre nasienie na roli swojej.</p>
<p>a – Matthew 13:24 – “Another parable He put to them, saying: ‘The kingdom of heaven is like a man who sowed good seed in his field’” – the texts are identical.</p>	
	<p>Heroda króla, oto mędrcy ze wschodu słońca przybyli do Jeruzalemu, mówiąc:</p>
<p>b – Matthew 2:1– “...Herod the king, behold, wise men from the East came to Jerusalem, saying:” – only in the facsimile, Jerusalem is uninflected in the genitive.</p>	
	<p>Tedy przystąpili do Jezusa z Jeruzalemu nauczani w Piśmie i Faryzeuszowie, mówiąc:</p>
<p>c – Matthew 15:1– “Then the scribes and Pharisees who were from Jerusalem came to Jesus, saying” – Jerusalem is inflected in the genitive in both texts. Also note the orthography &lt;y&gt; for modern &lt;i&gt; ‘and’. The discrepant &lt;a&gt; after the second word in the facsimile marks a note at the side of the page, and is not part of the text.</p>	
	<p>la; ale który jest najmniejszym w królestwie niebieskim, większy jest,</p>
<p>d – Matthew 11:11– “...but he who is least in the kingdom of heaven is greater....” – the neuter adjective &lt;niebieskim&gt; ‘heavenly’ is spelled &lt;niebieskiem&gt; in the 1881 edition.</p>	
<p>Fig. 2.1: The same text from a facsimile of an edition from 1632 next to the digital text of the Gdansk Bible. (scans from <a href="http://www.bibliagdanska.pl">http://www.bibliagdanska.pl</a>)</p>	

Since the older text is a Protestant Bible, the newer translation was taken from the contemporary Protestant Bible, and not from the more widespread Polish Catholic Bible, in order to avoid possible discrepancies stemming from underlying theological differences. The text therefore comes from the 1990 edition of the Warsaw Bible (Biblia Warszawska), first published in 1975, which finally replaced the archaic Bible of Gdansk as the standard Polish Protestant Bible (available e.g. at <http://www.bapost.ok.info.pl/nt/>). Since the translation work had access to and consulted the Gdansk Bible, the texts are generally similar, meaning that a good parallelism can be expected compared to other translations. It is however likely that the influence of the Gdansk Bible may have led to a

conservative adoption of its forms in certain places. As discussed in section 1.2, this can be viewed as a feature of biblical language (which is likely deeply influenced by previous translations in most languages), and at the same time, may give greater weight to those differences that are found between the texts in spite of this.

The entire parallel corpus with both texts contains a little over 46,000 tokens, in 1,071 aligned verses. The small size in terms of a normal, mono-lingual corpus is partly made necessary by the lack of reliable training data for tagging the older language, meaning annotation must be manually proofread. On the other hand, this also ensures high quality tagging, and the size has proven to be sufficient for the application of many statistical measures with satisfactory accuracy. It may be noted in this context that many parallel corpus-based techniques in machine translation often achieve various tasks at good success rates with well below 1,000 example pairs (Somers, 1999: 119-121). Furthermore, smaller corpora can provide very similar results to larger ones, provided they are likewise homogeneous (see Nurmi, 2002 for a positive evaluation of monolingual research with a larger historical corpus versus a smaller subset of it). For relatively frequent phenomena, and especially morphology and certain parts of the lexis (cf. chapters 3 and 4), and somewhat less so for syntax and grammatical categories (chapter 5), the interdependency between the two texts allows drawing founded conclusions from comparably little data, provided annotation quality is high and the parallelism is faithful. According to Fung (1998: 2), a successful extraction of correspondences from parallel corpora depends on the degree of conformity to the following characteristics:

- o Words have one sense per corpus.
- o Words have a single translation per corpus.
- o There are no missing translations in the target document.
- o The frequencies of words and their translations are comparable.
- o The positions of words and their translations are comparable.

These properties seem to generally hold with regard to the Bible text, which is typically translated very painstakingly and completely, and is also semantically relatively



homogeneous, reducing polysemy. The similarity of the (sub)language stages also contributes to similar word order and comparable frequencies. All these factors contribute to a good starting point for the extraction of correspondences using a parallel diachronic Bible corpus. In the following I will refer to the corpus taken from the Gdansk Bible as GMat or simply G, and to the newer Warsaw Bible corpus as WMat or W.

## 2.2 Tokenization

Polish orthography makes tokenizing a relatively easy task, and all the more so when dealing with such a regular text as the New Testament, which contains no abbreviations, numeric tokens, etc. I have generally followed the same principles used in the IPI PAN Corpus (Przepiórkowski, 2004), the largest electronic corpus of Polish on-line at the time of writing: tokens contain no whitespace (precluding multi-word “New York”-style tokens, which in any case do not seem to appear in this corpus), and either contain no punctuation, or only punctuation marks (cf. Przepiórkowski and Woliński, 2003). Tokens thus almost always correspond to strings of characters separated by spaces, or the common punctuation marks (e.g. ‘.’, ‘!’, etc.), which are tokenized separately from the adjoining words. Verses are identified by their running numbers which appear as numerals in the original text; these numerals are not themselves tokenized, but rather removed from the text stream altogether.

However, there are problematic cases for the space/punctuation-based tokenization rule, involving enclitic elements that are written together with whatever orthographic word precedes them. Some emphatic particles, for example, are enclitic (Swan, 2002: 410-411) and often appear after the first tonic unit in a sentence or clause, i.e. in the Wackernagel position (Wackernagel, 1882). Such a particle is shown in boldface in example (1)<sup>3</sup>. It might have been desirable to simply separate these tokens, treating their univertized orthography as insignificant, but there are also cases where the clitic changes the word form it is attached to, so that the separated form is no longer identical to any independent word form. In example (2), for instance, the normal form of the interrogative pronoun *co* ‘what’ is changed into *có* before the enclitic emphatic

---

<sup>3</sup> In examples throughout this work I use glosses which are as literal as possible to facilitate understanding, at the cost of an incomplete annotation where grammatical categories are not relevant for the question at hand. See appendix A for a list of abbreviations used in the glosses.



particle *ż*, as a result of a phonological process conditioned by the closed syllable produced by the presence of the clitic.

- (1) Gdzie<sup>ż</sup> jest ten, który się narodził, król żydowski?  
 where+**EMPH** is this which REFL was-born king Jewish  
*Where is the one who was born, the Jewish king?* (GMat 2:2)
- (2) có<sup>ż</sup> będziemy jeść?  
 what+**EMPH** we-will eat  
*what will we eat?* (GMat 6:31)

Furthermore, these fused units may be lexicalized and develop a special meaning of their own, which can be discerned from the appearance of a separate dictionary entry for them, e.g. *gdyż* ‘since, because’, comprised of the word *gdy* ‘when, if’ and the particle *ż*. For the older language, it is all the more difficult to judge which forms should be separated. Since it is impracticable to study each such case separately in the scope of this work, it was decided to use two alternative tokenizations in parallel. One tokenization follows the practice of the IPI PAN corpus: it is maximally fine-grained, separating clitics where possible, and normalizing changed forms through consistent lemmatization (i.e. the lemma of the token encompassing <có> is *co*). It relies on the dictionary used by the tagger (see next section) to determine clitic status: if a form containing a clitic is not in the dictionary, it is broken up. The other tokenization breaks up no orthographic strings, and since it only conflicts with the first tokenization in a few cases, it is accomplished using special additional link tokens, which have a property ‘links’ listing the items involved in the cliticization, and containing no word form. The tokens representing the sequence <có<sup>ż</sup>> in example (2), for example, are the following:

```
<t ID='g1c06v31s01t08' lemma='co' pos='ProIr' case='acc' num='sg' gend='N'>có</t>
<t ID='g1c06v31s01t08b' links='g1c06v31s01t08;g1c06v31s01t09'></t>
<t ID='g1c06v31s01t09' lemma='ż' pos='PtcI'>ż</t>
```

In some cases, the same elements appear in GMat with a clitic and in the WMat without it, and the similarity between the texts can then be recognized only in the fine-

grained tokenization. For instance, examples (3) and (4) have the parallel text to examples (1) and (2), which have the same interrogative pronouns without the emphatic *ż*.

- (3) Gdzie jest nowo narodzony król żydowski?  
 where is newly born king Jewish  
*Where is the newly born Jewish king?* (WMat 2:2)
- (4) co będziemy jeść?  
 what we-will eat  
*what will we eat?* (WMat 6:31)

A more complicated challenge is presented by vestigial forms of the enclitic auxiliary verb ‘to be’ which are written together with the preceding word. Mostly these clitics appear in the past tense following a perfect participle, e.g. *widzieliśmy* ‘we saw’ = *widzieli* ‘see (perfect participle)’ + *śmy* ‘we are (auxiliary)’, where they are often interpreted as inflectional endings (e.g. Swan, 2002). This is the result of a univerbized periphrastic construction which began fusing as early as the 15<sup>th</sup> century (Rospond, 2003: 178-179). But in some cases the clitic may still ‘separate’ from its participle even in Modern Polish, appearing instead in the ‘second’, Wackernagel position (see Swan, 2002: 255). Compare example (5) (fused auxiliary) with (6) (detached auxiliary):

- (5) Chleba nie wzięliśmy  
 bread not take-PARTPF+AUX-1-pl  
*We did not take bread* (WMat 16:7)
- (6) bośmy widzieli gwiazdę jego  
 for+AUX-1-pl see-PARTPF star his  
*for we have seen his star* (GMat 2:2)

Another common case involves conditionals, which use the same participle and clitics, the latter usually attached to a conjunction like *gdyby* ‘if’, *aby* ‘so that’, and a few others:

- (7) Gdybyśmy żyli za dni ojców naszych  
 if+AUX-1-pl live-PARTPF during days fathers’ our  
*If we had lived in the days of our fathers* (WMat 23:30)

**Fig. 2.2: Tokens for the text in example (6).** The last token annotates a past tense periphrasis with the ‘links’ property, allowing the search engine to display an annotation spanning the relevant items.

Since it may be interesting to look at past tense or conditional verbs in one context, and at the position of clitics in another, both notations are made available. The clitic is tokenized separately, allowing lemmatization of the separate lexemes. At the same time, the entire construction, expressing e.g. a past tense verb and ignoring the detached clitic, is annotated separately in a further token holding no word form, with properties indicating the lemma and inflection of the construction (person, number etc.), as well as the tokens involved in the periphrasis and enclisis. It is thus possible to search for either past tense verbs or clitics, and, since the items involved in the cliticization are marked up using the ‘links’ property, to represent these analyses graphically in a search engine. Figure 2.2 illustrates the notation of the tokens in (6) and their presentation in a browser-based search engine constructed to display these records. The engine provides basic search functions on the annotation of single tokens or adjacent fixed length token sequences, and a parallel verse KWIC display. Since it is immaterial to the rest of this work, it will not be described further at this point.

Clearly, quantitative results will be affected by the choice of tokenization used, but this is not necessarily bad: the annotations are consciously different interpretations of the same data (cf. Lüdeling, 2007). For the most part, I will use the fine-grained annotation, which is useful for comparing occurrences of lemmas in the corpora (e.g. parallel *co* and *có-ž* will register a cooccurring lemma). The second tokenization will be preferred only in chapter 5, which is concerned with grammatical functions, in order to abstract over grammatical categories (e.g. all occurrences of the past tense).

## 2.3 Morphophonological Tagging with Polimorph

The corpus was tagged and lemmatized using a tagging program called Polimorph (Zeldes, 2006) and a digital version of Swan's *A Learner's Polish-English Dictionary* (available from the author's web site at <http://polish.slavic.pitt.edu>), which was enriched with some entries to cover missing older lemmas. Since the orthography of the electronic Gdańsk Bible was modernized, it was generally possible to use the modern dictionary for both texts. For morphological disambiguation it was found to be very useful to use annotation projection: WMat was tagged first, and scores for GMat tags were boosted if they matched lemmas and tags in parallel sections of WMat<sup>4</sup>. A possible bias in GMat's tagging to wrongly resemble WMat's was prevented by a complete manual proofreading of both tagged corpora, which was in any case helpful to ensure an accurate tagging.

The Polimorph tagger uses a generative phonological model, deriving a variety of allomorphic surface forms of Polish morphological suffixes from one underlying form per suffix. An advantage of this approach is that the tagger can output a linguistically motivated phonological representation of the suffix used to analyze and identify each form and its lemma, and these can be annotated in the corpus (this will be taken advantage of in chapter 3). This means, for example, that suffixes containing different allophones of the same phoneme follow a uniform notation for the underlying phoneme, e.g. /y/ for both allophones <i> = [i] and <y> = [y] in the suffix of the nominative singular masculine adjective: <ciężki> 'heavy (nom. sg. masc.)' with the <i> suffix conditioned by the preceding velar, but <piękny> 'beautiful' with <y>.

Additionally, the suffixes follow a morphophonological notation along the lines used in Swan's (2002) grammar. This means that suffixes are also analyzed in terms of the morphophonological alternations (or 'mutations') that they cause in the stems to which they are attached, which is relevant especially for distinguishing some otherwise homographic suffixes. For example, two different suffixes containing the phoneme /y/ mark the aforementioned form <ciężki> 'heavy (nom. sg. masc.)' and one of its plural forms <ciężcy> 'heavy (nom. pl. masc. personal)'. Although both suffixes consist of the phoneme /y/, the first suffix palatalizes the stem's final /k/ into a palatovelar /k'/, while

---

<sup>4</sup> This is facilitated by the high amount of lexical overlap between the texts, on which see chapter 4. Failing such overlap, more sophisticated projection techniques would be needed (cf. Yarowski and Ngai, 2001).

the second mutates the /k/ into an affricate /c/ ([ts]). In order to represent this difference, the first suffix is notated as R4y# and the second is notated as R1y#, where the capital R followed by a number (between 1 and 4) represents a ‘mutation operator’ (cf. Swan, 2002: 23-26), indicating which of four characteristic mutation types the suffix may cause in the stem it is attached to. In this work I will only be interested in using these operators to distinguish between homographic suffixes; for a complete account of the four mutation operators see Swan (2002: 23-42) and Zeldes (2006).

Crucially for the study of morphology in chapter 3, the lexicon used by Polimorph does not specify which suffixes may appear with which lemma. This means that non-standard, fluctuating or archaic forms are all accepted as possible beside regular forms, regardless of which is standard. This is similar to an English tagger accepting the normative form <oxen>, with the irregular suffix <en>, as a plural of the lemma *ox*, as well as a possible regularized <oxes>. In this way the tagger does not smooth out morphological variation in the corpus by using a normative suffix list – any possible combination of lemma and suffix is accepted, as long as constraints regarding part-of-speech, gender, etc. are respected (for more details see *ibid.*).

## 2.4 The Tag-Set

In preparing the tag-set used in this corpus, the so-called ‘flexemic’ tag-set (Woliński and Przepiórkowski, 2001; Przepiórkowski and Woliński, 2003) used in the IPI PAN Corpus (Przepiórkowski, 2004), was taken as a starting point. The guiding principle behind this tag-set is morphological and morphosyntactic, meaning that it classifies words according to which variable morphological features they exhibit (e.g. case for nouns), and which lexical ones (e.g. gender for nouns, which do not inflect for gender, but agree with adjectives on this property morphosyntactically). The tag-set has been represented by its authors as a decision tree, where binary properties such as “inflects for gender” or “has person” are used to break down candidates into finer classes. At the lowest level, some distinctions based on closed lists and orthographic properties are also used where morphological criteria could not produce semantically desirable classes (e.g. the list of prepositions, or the property “ends in -no or -to” to distinguish the indeclinable impersonal past form from infinitives and other indeclinable adverbial participles).

The tag-set described below departs from the IPI PAN tag-set on several points, which were deemed necessary or preferable for this study. Wherever possible, it has been attempted to allow the possibility of an easy reversion to the IPI PAN tag-set by providing tag names that can be equated with the IPI PAN tags by selecting a substring of the tag name. As one example, the distinction between common nouns and proper nouns, which is ignored in the IPI PAN corpus, has been upheld in this corpus. Notwithstanding theoretical debate about special properties of proper names that may warrant their separate tagging, there are some practical reasons for this. One reason already noted in section 2.1 is that proper nouns exhibit greater divergence with the print edition of the older text and are thus more unreliable for morphological study. Additionally, the high proportion of non-nativized Hebrew names (e.g. irregular feminines like *Tamar*, as opposed to nativized names like *Piotr*) skews the distribution of suffixes in favor of some otherwise uncommon inflectional patterns. It is therefore desirable to be able to consider only common nouns. To allow both separate and joint queries, the tag S is assigned to common nouns and SN to proper nouns. A query for tags starting with S returns all nouns.

Table 2.1 describes the tags used in the corpus, alongside their IPI PAN counterparts and examples. The table reveals much similarity between the tags, with the Polimath tags usually refining the IPI PAN tags. The most central difference between the tag-sets is the inclusion of a tag for finite verbs, VFin. Since past tense forms are originally based on periphrastic constructions with participles (cf. section 2.2) they inflect for gender, placing them in a different flexemic class according to Przepiórkowski and Woliński (2003). At the same time, there is no representation of the unverbized participle and auxiliary in the IPI PAN corpus (the second tokenization discussed in section 2.2), meaning that one cannot query for only past tense forms, or only conditionals – one can only look for the perfect participle, in whatever use. With a view to investigating correlations between non-finite and finite constructions between the corpora, I decided to define the VFin tag, and allow it to receive a gender property in the past tense, missing from the other tenses<sup>5</sup>. Since the conditional mood is also tagged as a type of VFin in this scheme, it has been thought sensible to also tag imperatives under

---

<sup>5</sup> Viewing these originally periphrastic forms synchronically as unverbized finite verbs is not the orthodox approach, but see Swan (2002: 245-255) for such a treatment; see also section 5.2 for one application of it.

VFin, using one property for both tense and mood (past, present, future, imperative or conditional), since these are all mutually exclusive and morphologically well defined.

Polimatth tag	Categories covered	Lemma form	Example (lemma)	IPI PAN
Adj	adjective	nom sg m positive	<i>polski</i>	<i>adj</i>
AdjComp				
AdjSuper		nom sg m (noun)	<i>dawidowy (Dawid)</i>	
AdjPos				
AdjPred	predicative adj.	sg masc. form	<i>powinien</i>	<i>winien</i>
Adv	adverb	positive form	<i>dobrze</i>	<i>adv</i>
AdvComp				
AdvSuper				
Cj	conjunction	same as form	<i>oraz</i>	<i>conj</i>
Prep	preposition	same as form	<i>na</i>	<i>prep</i>
ProDm	demonstrative	nom sg m	<i>ten</i>	<i>(adj)</i>
ProIr	interrogative	nom sg	<i>kto, co</i>	<i>(subst)</i>
ProNm1	numeral	nom sg m	<i>jeden</i>	<i>(adj)</i>
ProNm2		nom pl m inanim.	<i>dwa</i>	<i>num</i>
ProNm34		nom pl m inanim.	<i>trzy</i>	
ProNmQ	collective numeral	nom pl m inanim.	<i>pięć</i> <i>wiele</i>	<i>numcol</i>
ProPr	personal pronoun	nom sg	<i>ja</i>	<i>ppron12</i>
		nom sg m	<i>on</i>	<i>ppron3</i>
ProPs	possessive pronoun	nom sg m	<i>mój</i>	<i>(adj)</i>
ProRf	reflexive pronoun	the form <i>siebie</i>	<i>siebie</i>	<i>siebie</i>
ProRl	relative pronoun	nom sg m	<i>który</i>	<i>(adj)</i>
Ptcl	particle	same as form	<i>nie, -że, się</i>	<i>qub</i>
Punct	punctuation	same as form	<i>.</i>	<i>interp</i>
S	noun	nom sg	<i>profesor</i>	<i>subst</i>
SN	proper noun	nom sg	<i>Jezus</i>	
SV	verbal noun	infinitive	<i>czytanie (czytać)</i>	<i>ger</i>
VAux	Auxiliary <i>być</i>	infinitive	<i>śmy (być)</i>	<i>aglt</i>
Vbd	predicative	same as form	<i>warto</i>	<i>pred</i>
VFin	non-past form	infinitive	<i>czytam (czytać)</i>	<i>fin</i>
	future <i>być</i>		<i>będzie (być)</i>	<i>bedzie</i>
	imperative form		<i>czytaj (czytać)</i>	<i>impt</i>
	past tense form		<i>czytałem/-m czytał (czytać)</i>	
	conditional form		<i>czytał bym czytał (czytać)</i>	
VConv	contemp.adv. part.	infinitive	<i>czytając (czytać)</i>	<i>pcon</i>
VInf	infinitive	infinitive	<i>czytać</i>	<i>inf</i>
VPartImpersPred	impersonal	infinitive	<i>czytano (czytać)</i>	<i>imps</i>
VPartPastAct	anterior adv. part.	infinitive	<i>przeczytawszy (przeczytać)</i>	<i>pant</i>
VPartPastPass	pass. adj. part.	infinitive	<i>czytany (czytać)</i>	<i>ppas</i>
VPartPf	l-participle	infinitive	<i>czytał (czytać)</i>	<i>praet</i>
VPartPresAct	act. adj. participle	infinitive	<i>czytający (czytać)</i>	<i>pact</i>

Tab. 2.1: Polimatth and IPI PAN tags.



Three further differences are the inclusion of more pronominal tags, sub-types of adjectives and adverbs, and a rearrangement of numeral tags. The demonstrative, possessive and relative pronouns are grayed out in the IPI PAN column, since they are treated as adjectives, and likewise the interrogative pronouns which are treated as substantives, all based on their morphological properties (though in fact interrogatives, unlike substantives, do not inflect for number, and may have variable gender in the case of ‘who’). However to investigate parallel constructions containing these categories one must be able to distinguish them from other adjectives and nouns, as they will probably have distinct parallels from these across the corpora, as well as distinct morphological behavior. Similarly, the distinctions of adjective degrees (positive, comparative and superlative) within the basic tag, instead of a “degree” property, was motivated by the possibility to explore their corresponding suffixal morphologies separately: if adjectives and comparatives have the same tag, a distribution of morphological affixes against part-of-speech will reveal multiple sets of adjectival or adverbial suffixes in each category. Likewise possessive adjectives, which are derived from nominal lemmas (i.e. words like *Jakubowy* ‘Jacobean’ with the sense ‘of Jacob / belonging to Jacob’), not only have their own suffixes and, unlike other adjectives, nominal lemmas, but can also be studied in their own right insofar as they have distinct correspondences within the diachronic corpus (cf. section 5.3). Nonetheless, an IPI PAN style search for all adjectives can be achieved using a wildcard query for tags beginning with Adj. Finally, the division of numerals in the IPI PAN tag-set makes no distinction between the numbers two and three/four, despite their different inflection: *trzy* ‘three’ and *cztery* ‘four’ distinguish only masculine personal versus non-masculine personal, whereas *dwa* ‘two’ also distinguishes feminine from neuter and masculine non-personal (Swan, 2002: 191). These are tagged differently in line with the flexemic principle, while collective numerals governing the genitive (e.g. *wiele* ‘many, much’) are grouped together with the cardinals above five, which behave in the same way morphologically and morphosyntactically. The numeral *jeden* ‘one’, which is tagged as an adjective in IPI PAN, is given its own tag on semantic grounds. Thus, most constellations of queries can be satisfied, and a search for all tags beginning with ‘ProNm’ delivers the ‘intuitive’ class of all numerals and related determiners of quantity.



The possible attributes of each tag and their values are given below:

**asp:** {*pfv*, *impfv*, *indet*, *det*, *freq*}. The aspect property characterizes all forms derived from verbal lemmas (tags containing V), including verbal nouns (SV), participles etc. In addition to the basic perfective and imperfective designations, the last three values represent sub-types of imperfective verbs, and are hence implicitly imperfective. Although the distinction between them is ignored in IPI PAN, they are commonly listed in dictionaries along with other aspectual partners, and so they have all been annotated; it is easy to ignore these values if this is desired. Only unprefixated motion verbs can be indeterminate or determinate, and they have no neutral imperfective counterpart, while frequentative verbs are derived from ‘neutral’ unprefixated imperfective verbs (cf. Swan, 2002: 290-293).

**gend:** {*F*, *N*, *M*, *MI*, *MA*, *MP*, *nV*, *V*}. Polish distinguishes neuter (N), feminine (F) and three types of masculine gender in nouns, most pronouns and adjectives: personal masculine (MP, denoting adult human males), animate masculine (MA, denoting masculine animals and, exceptionally, some inanimate substantives) and inanimate masculine (MI). The value masculine (M) is used ambiguously in adjectives, implying agreement with any masculine lemma. Virile (V) and non-virile (nV) forms characterize plural adjectives, participles and pronouns agreeing with groups containing at least one masculine personal referent, or no such referent respectively<sup>6</sup>. In contrast to IPI PAN conventions, additional genders based on different types of plural only nouns (nouns that have no singular form, like *drzwi* ‘door/doors’) are not recognized<sup>7</sup>. These forms can however be identified and retrieved by looking for lemma suffixes (see below) characteristic of plural forms.

**case:** {*nom*, *gen*, *dat*, *acc*, *inst*, *loc*, *voc*}. The seven grammatical cases in Polish, characterizing nominal, adjectival and most pronominal forms.

---

<sup>6</sup> The term virile is used by some interchangeably with the MP gender. Although all plural MP nouns have virile congruence, not all virile adjectives refer exclusively to MP pluralities – they can refer to mixed groups. I use ‘V’ and ‘nV’ specifically to refer to the two way opposition in non-lexical gender of variable gender elements such as adjectives, in the plural.

<sup>7</sup> In fact, Adam Przepiórkowski, one of the authors of the IPI PAN tag-set, has also opposed the use of these genders himself (Przepiórkowski, 2003), reverting to Witold Mańczak’s (1956) classic view of five nominal genders.

**num:** {*sg, pl*}. Although some traces of its inflection remain in the older language, the dual number is not annotated separately from the plural since there is no formal dual congruence with adjectives, pronouns, etc.

**pers:** {*1, 2, 3*}. First, second and third person of finite verbs and personal pronouns. Note that impersonal verb forms (Vbd and VPartImpersPred) have no value for this property.

**tense:** {*past, pres, fut, imp, cond*}. As already discussed above, the ‘tense’ attribute doubles to encompass mood as well: imperative, conditional, or in the case of the other three tenses, implicitly indicative. This is possible since Polish modal forms are not marked for tense, and vice versa.

**suf, lsuf:** [*string*]. Any inflected form may be characterized by a suffix and a lemma suffix, as arrived at during morphological analysis. If a token exhibits a suffix but no lemma suffix, then its lemma is suppletive (i.e. there is no direct connection between the inflected form and the lemma, as in *lepsz* ‘better’ and its lemma *dobry* ‘good’, which have different, unrelated stems). If an inflected form shows neither suffix, then its form is part of a closed class of irregularities (e.g. *ja* ‘I’ has no identifiable suffix).

### 3 Morphological Suffix Change

A parallel diachronic corpus with morphological suffix annotation provides two diachronically disparate, but interdependent samples of inflectional distributions. The purpose of this chapter is to find out whether it is possible to identify which Polish morphological suffixes<sup>8</sup> have changed and into which, relying solely on corpus data. In order to accomplish this, the fact that the language stages are closely related can be taken advantage of in two different ways. Firstly, the distributions of morphological suffixes can be expected to be more similar than in non parallel corpora by virtue of the corpora's identical content, meaning even small deviations in distribution based on relatively little data can be very meaningful. Secondly, since the orthography in GMat has been normalized and lemmatization has been carried out with a modern lexicon (cf. section 2.3), it can be expected that many cases will be found where the same lemmas are used in parallel with the same grammatical functions, but possibly with different suffixes. One may thus be able to identify changes in suffixal morphology by searching for minimal pairs of tokens with identical lemmas and grammatical analyses (case, gender, number, etc.), but different suffixes. Such pairs of tokens are made possible by the Polimorph tagger, which, as mentioned in section 2.3, uses a dictionary that does not specify the list of permissible suffixes for each lemma; instead, it accepts any suffix which may be used to create a regular form of any lemma as a possibility for analysis.

Though the approach outlined above can apply equally well to all inflected word forms, for space reasons, I will limit the discussion in this chapter to nominal inflection. I choose this area since it shows the most extensive morphological change in the period in question. The next section therefore gives some preliminary background remarks on the history of Polish nominal declensions. The following two sections continue with an exhaustive investigation of the multiple minimal suffix pairs that can be discovered in Polimatth, along with the quantitative distributions of these suffixes, and finally, of all nominal suffixes in an overview. The results will be discussed and evaluated in light of known treatments of Polish, and more generally, Slavic historical morphology.

---

<sup>8</sup> Suffixal morphology is given the center of attention here since it encompasses almost all of Polish inflectional morphology, if we disregard the superlative prefix *naj-*. Prefixed elements, which play a role in word formation, will be discussed in chapter 4.

### 3.1 Preliminary Remarks on Polish Declensions and their Origins

In order to understand the historical changes in Polish suffixal morphology, brief mention must be made of the origins of the different inflectional types one finds in Middle and Modern Polish. Polish is an Indo-European language, of the West-Slavic branch of the Slavic language family. As such it inherited 7 grammatical cases from the original 8 postulated for Proto Indo-European: nominative, vocative, accusative, dative, locative, instrumental and genitive. As in all Slavic languages, the latter case preserves the form of the old Indo-European ablative case in some paradigms (cf. Beekes, 1995: 190-192).

Noun inflectional types in Indo-European languages are often classified according to the last phoneme their stems are thought to have had in Proto Indo-European, before the inflectional ending for case, number and gender. The o-stems, which had the vowel \*o at the end of their stems, are the most common type of noun in most Indo-European languages, and account for the inflection of almost all Polish masculine and neuter nouns. The o-stem nominative and accusative endings in Polish are zero (i.e. a phonologically empty -Ø, or #<sup>9</sup>) as in *dół* ‘pit, bottom’; this comes from the disappearance of the Proto Slavic ending \*ŭ (cf. Old Church Slavonic *dolŭ* ‘hole, pit’) and ultimately from \*os in the nominative and \*om in the accusative of Proto Indo-European \*d<sup>h</sup>olo-s, \*d<sup>h</sup>olo-m (cf. English *dale*, German *Tal* ‘valley’ etc.). Most feminine nouns in Polish come from the a-stems (sometimes called h<sub>2</sub>-stems, since the phoneme /a/ at their end derives from a reconstructed laryngeal phoneme marked with the symbol h<sub>2</sub>). These nouns end in a# in Modern Polish as well, e.g.: *żona* ‘wife’.

Other, less frequent types in Polish are the feminine i-stems, which synchronically in Polish are feminine nouns that end in a consonant (e.g. *noc* ‘night’), and the non-productive neuter n- and nt-stems, which are neuters ending in the nasal vowel /ɛ/ (e.g. *imię* ‘name’). There are furthermore some individual case forms, both regular and irregular, that come from other classes which have otherwise gone out of use. The old u-stems especially have left alternate inflectional forms in the masculine declension, such as the u-stem dative in owi# beside the o-stem dative in u# (e.g. *chlebowi* ‘bread (dat.)’ but *psu* ‘dog (dat.)’). The choice of suffix no longer depends on which stem type a word

---

<sup>9</sup> In this chapter I will use the word border sign ‘#’ to denote the end of a suffix. ‘#’ alone therefore stands for the zero suffix.

once had – in Modern Polish, former o-stems can take u-stem endings and vice versa. Old stem types that are no longer in use are thus a source for alternative forms and possible morphological change.

Another important distinction in Polish nominal inflection is between so-called functionally hard stems, ending in one of the consonants /p/, /b/, /f/, /w/, /m/, /t/, /d/, /s/, /z/, /ʃ/, /r/, /n/, /k/, /g/, /ch/<sup>10</sup>, and functionally soft stems, ending in other consonants. Nouns have somewhat different endings in certain grammatical cases depending on whether their stems are soft or not, and stems ending in velars additionally have some further differences (see Swan, 2002: 23-24, 66). The soft/hard distinction originally stems from palatalization caused by a front vowel or glide in a suffix: e.g. o-stems ending in \*-io-s produced soft stems, while other o-stems ending in plain \*-o-s produced hard stems (see Bielfeldt, 1961: 118-120). We may thus find changes that only occurred in either soft or hard stem nouns, but also in both.

It is customary in historical Indo-European linguistics to speak of stems comprised of roots and suffixes, followed by inflectional endings. In the example above, \**dʰol-o-s*, the second /o/ is the suffix and /s/ is the nominative singular case ending. In Polish, this division is no longer possible: the Ø at the end of *dół* historically encompasses both suffix and ending, and similarly, in defunct u-stem endings like *owi*#, the <w> belongs to the old u-suffix, and the <i> to the case ending, and other case forms of the same noun will not show the <w> at all, since the entire morpheme *owi*# may have been carried over to a noun that was not originally a u-stem. I will therefore use the term ‘suffix’ to refer to whatever morpheme synchronically marks a case ending, regardless of its mixed origins. This is also the policy used by the tagger to strip suffixes and identify morphological forms (cf. section 2.3 and Zeldes, 2006), as well as the way these suffixes are commonly described in synchronic Polish grammars (e.g. Swan, 2002).

### 3.2 Nominal Suffix Changes in Minimal Pairs

Since the Polimatth corpus is in essence ‘flat’ (i.e. lacks hierarchical annotation, notwithstanding link tokens, cf. section 2.2), it can easily be represented in a database table. Retrieving lemmas with multiple suffixes for the same form can then be achieved

---

<sup>10</sup> The last phoneme is represented using its orthographic digraph <ch> for convenience. It stands for [x]/[ç].

using an SQL query. The following query groups lemmas and grammatical analyses in a unified table containing both corpora, and counts how many suffixes mark each of them. It then only retrieves those analyses having more than one suffix:

```
SELECT lemma, pos, asp, tense, pers, num, case, gend, Count(suf)
FROM (SELECT lemma, pos, asp, tense, pers, num, case, gend, suf
      FROM GMat UNION SELECT lemma, pos, asp, tense, pers, num, case,
      gend, suf FROM WMat)
WHERE not suf Is Null
GROUP BY lemma, pos, asp, tense, pers, num, case, gend
HAVING Count(suf) > 1;
```

The suffixes characterizing these forms can now be retrieved, grouped and arranged, e.g. by part of speech. Table 3.1 lists the results for all different suffix pairs in the parallel corpus which mark the same form of the same common noun lemma (part-of-speech tag S). For the analyses recall that Polish distinguishes three masculine genders: personal (MP), animate (MA) and inanimate (MI); M means any one of these (cf. section 2.4). All of the alternations in the table correspond to historical developments in Polish nominal morphology<sup>11</sup>. In the following subsections I will discuss the phenomena corresponding to these results as they are described by historical grammars, and explore their distributions in the corpora.

Row	Analysis	Suffix Pairs		Examples		Sense
1	acc pl MP	R4y#	ów#	<i>anioły</i>	<i>aniółów</i>	angels
		R4e#	ów#	<i>króle</i>	<i>królów</i>	kings
		R4e#	R4y#	<i>nauczyciele</i>	<i>nauczycieli</i>	teachers
		#	R4y#	<i>śług</i>	<i>ślugi</i>	slaves
2	gen pl MP	ów#	#	<i>poganów</i>	<i>pogan</i>	heathens
3	gen sg MI	a#	u#	<i>podółka</i>	<i>podółku</i>	hem
4	inst pl M/N	R4y#	ami#	<i>duchy</i>	<i> duchami</i>	spirits
5	inst pl MI	mi#	ami#	<i>kijmi</i>	<i>kijami</i>	clubs
6	inst pl N	R4yma#	ami#	<i>uszyrna</i>	<i>uszami</i>	ears
7	nom pl MP	owie#	R4e#	<i>wężowie</i>	<i>węże</i>	snakes
		owie#	R4y#	<i>narodowie</i>	<i>narody</i>	peoples
8	nom/acc pl F	R4y#	R4e#	<i>noCy</i>	<i>noce</i>	nights
9	acc sg MP	#	a#	<i>(wyjść za) mąż</i>	<i>męża</i>	husband

Tab. 3.1: Variant suffix pairs in nominal morphology.

<sup>11</sup> However, the last entry, due solely to the expression *wyjść za mąż* ‘to marry (a man)’ (lit. ‘to go out behind a husband’), contains a fossilized accusative *mąż* ‘husband’ with an old zero suffix, which was quite possibly no longer transparent already in the 17<sup>th</sup> century. The frozen minimal pair *mąż* : *męża* ‘husband’ in fact attests a change that occurred already in Old Polish, and not internally between the periods represented in these corpora.

### 3.2.1 Masculine Personal Plural

Many of the changes in nominal morphology revolve around the evolution of the masculine personal (MP) gender, a gender reserved for nouns denoting adult male human beings. This gender not only has some different suffixes, but a distinct form of congruence, sometimes referred to as ‘virile’ (V), with adjectives and pronouns in the accusative plural: masculine personal nouns and their congruent attributes take the same form for the accusative plural as for the genitive plural, as opposed to the non-virile (nV) plural, which includes feminine, neuter, and non-human (but possibly animate) masculine nouns, for which the accusative plural is identical to the nominative plural. This distinction creates a binary division V : nV in the plural, which, along with the singular distinction between feminine (F), neuter (N), masculine inanimate (MI) and masculine animate (MA) nouns, forms the basic five-gender system of Polish (see Mańczak, 1956 and Przepiórkowski, 2003).

The distinction has its roots in the Common Slavic<sup>12</sup> syncretism of the nominative and accusative masculine singular of the o-stems, which led to the use of the genitive form instead of the accusative singular, in order to distinguish subject and object in sentences with two animate masculine participants (probably by analogy to the negative<sup>13</sup> and partitive genitives, or simply to verbs taking a genitive object; for an overview see Vaillant, 1977: vol. 5, 37-50). Like other Slavic languages, Polish later came to replace the accusative plural (which was still morphologically distinct) with the genitive when referring to animate masculines by analogy to the singular. However, in Polish this was done only when the referents were human, thereby creating the further subdivision into inanimate masculines (where nom.=acc.), animates (where acc.=gen. in sg. but nom.=acc. in pl.) and viriles (acc.=gen.). This analogy first took place in pronouns and then in nouns (see Klemensiewicz et al., 1955: 271-272, 281-282).

The first row in Table 3.1 above shows precisely this replacement of the old nominal accusative plural forms with the new analogical genitive forms: R4y#, R4e# >

---

<sup>12</sup> Common Slavic is the term used to refer to the last stage of Slavic linguistic unity, before the division into the East, West and South Slavic branches.

<sup>13</sup> The negative genitive, which is attested in the oldest Slavic monuments, replaces the accusative as the direct object of a negated verb, as in Modern Polish: *widziałem psa* ‘I saw a dog (acc.)’ but *nie widziałem psa* ‘I didn’t see a dog (gen.)’.

ów# and R4e# > R4y# (note that R4y# consequently still marks some accusatives, but for the nouns where it does so, it is in fact the gen. form, while their original acc. was R4e#).

Similarly, so called common-gender nouns (also called ‘epicenes’), which are masculine nouns that had feminine morphology (ending in the nominative with the feminine suffix a#) but could also designate male humans, sometimes replaced the feminine accusative plural with the corresponding feminine genitive form (which has a ‘zero’ suffix, #), creating such pairs as R4y# : # in *ślugi* : *śluga* (from *śluga* ‘servant (masc. and fem.)’). When the latter form appears in GMat, any adjectives or pronouns qualifying it appear in the genitive as well, instead of the expected accusative masculine plural:

- (8) posłał ślugi swoje  
sent servants his  
*he sent his servants* (GMat 21:34, *servants* has the old acc. fem. form)

but also:

- (9) posłał inszych ślug  
sent other servants  
*he sent other servants* (GMat 21:36, *servants* has the gen.=acc. form)

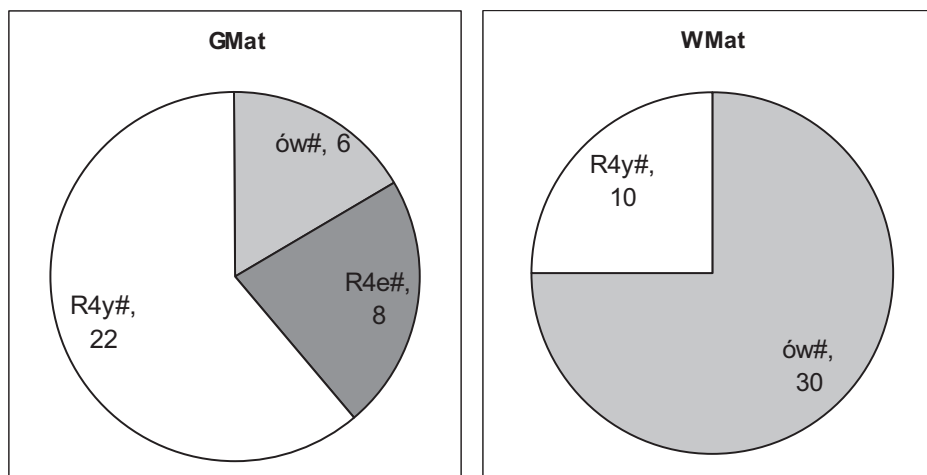
Most such nouns nowadays take the masculine genitive suffix ów# for the genitive and accusative plural, showing a harmonization of morphology and grammatical categories, but the archaic word *śluga* in particular normatively keeps the old feminine accusative *ślugi* (Swan, 2002: 84, 101).

Another analogical change which was not fully accepted is the genitive plural in ów#, originally a u-stem ending, for nouns in *-anin*, which usually designate nationalities or ethnonyms. These nouns drop the stem extension *-in-* in the plural, creating pairs such as *poganin*, gen. pl. *pogan* ‘pagan’. As shown in row 2 of Table 3.1, GMat shows analogical genitives in ów# (*poganów*), instead of the older normative gen. pl. in #, which is the current form for most nouns. Some exceptions in ów# do however exist in the contemporary language, e.g. *Cyganin* ‘gypsy’, gen./acc. pl. *Cyganów* (Swan, 2002: 98).

The two graphs in Figure 3.1 show token counts for each of the masculine personal accusative plural suffixes in the two corpora (excluding the feminine-like



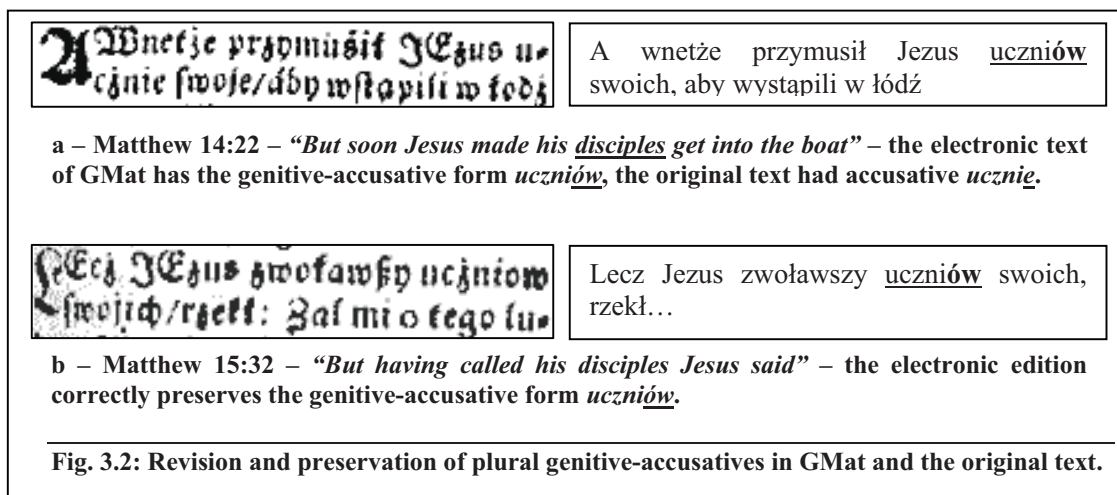
epicenes). If the information in Table 3.1 is not considered, these graphs are misleading. R4y# represents an old accusative suffix in GMat, but in WMat it historically corresponds to a genitive suffix (as mentioned above, in nouns where the old accusative was R4e#). Consequently the WMat graph is essentially showing the distribution of genitive suffixes serving as accusatives, while the GMat one shows two dominant conservative accusative suffixes (R4y#, R4e#), and the introduction of the new genitive-accusative plural (the suffix ów#).



**Fig. 3.1: Accusative masculine personal plural suffixes in both corpora.**

Thus, although distribution queries can show that ów# has become the dominant suffix for the accusative plural of MP nouns, and that R4e# has completely died out, they cannot distinguish that R4y# in WMat has a different source than R4y# in GMat. This fact can only be revealed by considering the suffix data from parallel lemma types in Table 3.1, although its significance and the reasons behind it can only be interpreted using corpus-external knowledge about the whole paradigm of each nominal inflectional class.

An additional problem I was able to discover with the GMat distribution, is that two of the cases of ów# seem to have been inserted, either knowingly or accidentally, by the 1881 revision. One of them appears in the 1632 print facsimile with R4e# (Figure 3.2a), and another with R4y#. The remaining examples were however verified in the facsimile (e.g. Figure 3.2b), showing the phenomenon of an old ów# acc. pl. to be even



more rare ( $4/36 = 11.1\%$ ), but correctly identified and attested already in the original contemporary text.

### 3.2.2 Masculine Inanimate Genitives: u# versus a#

The fluctuation of the genitive masculine singular between the suffixes a# and u# on row 3 of Table 3.1 is part of a known trend to make animate masculine nouns have the genitive in a#, and inanimates in u# (Rospond, 2003: 126-127). This process is still ongoing in contemporary Polish, with endings changing in both directions, though considerable groups of exceptions persist (Swan, 2002: 72-73). It should be noted that in this case, both tokens in the example come from GMat (since the query recovered elements with multiple suffixes in a union table of both corpora, it also recovers this fluctuation within one corpus); all the more remarkable considering this fluctuation was left in tact by the revisions of the text (the readings are confirmed by the facsimile). Since we are not fortunate enough to find a pair with two forms across the corpora (though a larger corpus might be expected to produce diachronic examples), the distributions of these suffixes can only be compared without regard to identical lemmas.

The graphs in Figure 3.3 give only a weak indication of this phenomenon – the distribution of the suffixes is similar in both texts. GMat shows a majority of a# genitives (upper white sections), but a higher frequency of u# genitives in the inanimate masculine (the two sections labelled MI), thus already in accordance with the known tendency to equate u# genitives with inanimacy. WMat shows much the same distribution, with a slightly larger proportion of inanimate u# genitives (110:70 or 61.1% instead of 92:76 or

54.7%), thus the tendency is, if at all, in the direction also known from the literature above. Nonetheless, inanimate *a#* genitives remain quite wide-spread. The other *a#* genitives form two groups, the large group of genitives marking persons (MP) and the small group marking animate genitives (MA), both of which are regular and expected.

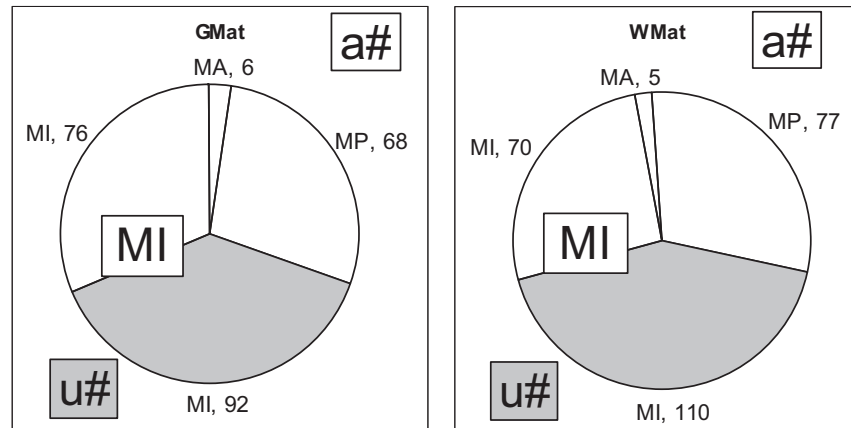


Fig. 3.3: Distributions of masculine genitive singular suffixes.

### 3.2.3 Instrumental Plural Masculine and Neuter

A clearer change can be seen in the instrumental plural suffixes of the neuter and masculine genders (rows 4-6 in Table 3.1). Here we see the loss of the old dual form (row 6), the irregular suffix *mi#* being thematized<sup>14</sup> into the regular suffix *ami#* from the pronominal and feminine nominal declensions (row 5), and the replacement of the old regular suffix for inst. pl. masc. and neut., which was *R4y#*, also by the new *ami#* (row 4). The spread of *ami#* at the expense of the other suffixes occurred over the course of the 17<sup>th</sup> century (Wiśniewska, 1994: 110-111), as reflected by the different corpus distributions in Figure 3.4.

The graphs confirm the disappearance of the old dual suffixes *oma#* and *R2yma#*, which appear only on the left. In Table 3.1 only *R2yma#* is found, since *oma#*, the hard stem allomorph of soft *R2yma#* (on soft and hard stems cf. section 3.1), has no parallel form with the same lemma and grammatical function. Since *oma#* appears only once in GMat, it is not surprising this correspondence is unattested, though it could certainly be expected, exactly as in the case of *R2yma#*, in a larger corpus. The last productive days

<sup>14</sup> i.e. given a vocalic onset. The term ‘thematic’ refers to suffixes beginning with a vowel. Athematic suffixes begin with a consonant, often creating assimilations in stem final consonants they are attached to.

of these two dual suffixes were probably in the 16<sup>th</sup> century, but use in nouns signifying natural duals such as hands, eyes etc. was still the norm well into the 18<sup>th</sup> century (Klemensiewicz, 1999: 304). The forms are now considered archaic (Swan, 2002: 119). It may also be noted that other dual case forms also happen to be unattested in Table 3.1 (there were a total of three distinct forms: nom./acc./voc., gen./loc. and dat./inst.); the gen./loc. form happens to be completely unattested in the corpus while nom./acc. dual appears only in the feminine *ręce* ‘hands’, which is however also the petrified plural of the word for ‘hand’ in Modern Polish, and thus remained unchanged.

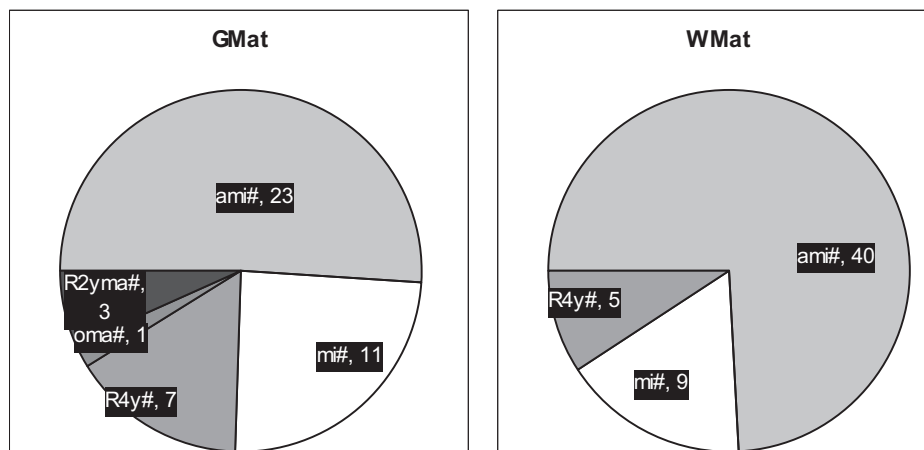


Fig. 3.4: Distributions of masculine and neuter instrumental plural suffixes.

Otherwise the graphs in Figure 3.4 show only a slight decrease in mi# and R4y#. However, an examination of the actual instances of the old R4y# in WMat shows it to be limited to a petrified use in the fixed expression *tymi słowy* ‘with these words’, appearing beside regular inst. pl. *słowami* ‘words (inst.)’ with the new suffix in productive use; in other words the old suffix was only retained where it was lexicalized (Wiśniewska, 1994: 110). Although it is possible to detect changes between the corpora automatically and also quantify them to an extent, attention must therefore always be given to the underlying data, which must be examined in order to ensure no artifacts are being produced by additional factors. The general trend is nonetheless quite clear: the analogical ami# suffix is gaining ground at the expense of all other suffixes and the disappearance of the isolated dual forms, growing from ≈51% to ≈74% in WMat, or even to ≈82% if the fixed expression is disregarded.

### 3.2.4 Nominative Plural Masculine

In row 7 of Table 3.1 we find a substitution of some nominative plurals in *owie#*, originally a u-stem ending (i.e. forming the nominative plural only of nouns whose stems had once ended in \*u, cf. section 3.1 above), with regular endings from the o-stems: *R4e#* (for soft stems) and *R4y#* (for hard stems). The new endings are probably motivated by a restriction of the *owie#* ending to substantives denoting male humans (strongly supported by the common u-stem *syn* ‘son’, which is frequent, and has the etymologically ‘correct’ nominative plural *synowie*). Words not denoting male human beings, such as *wąż* ‘snake’ and *naród* ‘(a) people’ (which in itself denotes many people, but is not a singular male human substantive) were thus given the more common o-stem plurals: *narody* with *R4y#* (a form which already appears in GMat beside *narodowie*, but exclusively in WMat) and *węże* with *R4e#*. Table 3.2 gives the result of a query for lemma types showing the *owie#* suffix in both corpora; attested forms are marked with *x*:

Form	Lemma	G <i>owie</i>	W <i>owie</i>	W <i>R4e/R4y</i>	English
aniołowie	anioł	x	x		angel
Chrystusowie	Chrystus	x			christ
królowie	król	x	x		king
mężowie	mąż	x	x		husband, man
świadkowie	świadek	x			witness
synowie	syn	x	x		son
uczniowie	uczeń	x	x		student
wodzowie	wódz	x			leader
budowniczo	budowniczy		x		architect, builder
wężowie	wąż	x		x	snake
narodowie	naród	x		x	people, nation
faryzeusze	faryzeusz	x		x	Pharisee
saduceusze	saduceusz	x		x	Sadducee

Tab. 3.2: Nominative plural masculine nouns showing the *owie#* suffix.

In all cases where a lemma appears with *owie#* in WMat, it represents a male human being-like entity, to the inclusion of ‘angel’ and ‘christ’. This corresponds to the main categories of *owie#* plural in Modern Polish as described by Swan: names of “male relations”, “deities and august rulers” and some ranks, titles and professions, especially when “used honorifically” (Swan, 2002: 79-80). The bottom four highlighted rows represent the item types whose plural forms have been changed in WMat. Among them are the two aforementioned cases which violated the ‘male human’ constraint, viz.

‘snake’ and ‘people’. The remaining two changes, ‘Pharisee’ and ‘Sadducee’, may well be due to the honorific force felt to be associated with *owie#* (as in Swan’s category “deities and august rulers” and the “honorific” use, cf. also the items ‘angel’, ‘christ’ and ‘king’ in the table). Thus the Pharisees and Sadducees, which are generally seen as very negative in the New Testament, may have been excluded from this privileged group.

### 3.2.5 Analogical Nominative-Accusative Plural Feminine i-Stems

Three pairs attest to a substitution of the suffix for both nom. and acc. feminine plural from *R4y#* to *R4e#* (row 8 in Table 3.1), all from the i-stems: *twarzy* : *twarze* ‘faces’, *mocy* : *moce* ‘powers’ and *niemocy* : *niemoce* ‘impotences, ailments’ (the previous word negated with *nie-*). The original ending *R4y#*<sup>15</sup> was replaced by *R4e#* in these and other feminines ending in soft consonants. This change occurred by analogy to the a-stems, the largest, most productive group of feminines, which have *R4y#* in the nom./acc. plural for hard stems, and *R4e#* for soft stems (Klemensiewicz et al., 1955: 298). In Modern Polish many such nouns take *R4e#*, some take the older *R4y#* (especially the large group of abstract nouns derived with the suffix *ość#*), and a few may still appear with either one (see *ibid.* and Swan, 2002: 46). The claim that forms with retained *R4y#* belong to common lemmas (e.g. in Swan, 2002: 46), while possibly true for Modern Polish in general, cannot be substantiated in WMat, as token frequencies of these lemmas show (i-stem forms have been isolated here by looking for feminines with a # lemma suffix):

Lemma	pos	num	gend	suf	lemsuf	f(lemma) in W	English
kradzież	S	pl	F	R4e#	#	1	theft
twarz	S	pl	F	R4e#	#	3	face
niemoc	S	pl	F	R4e#	#	4	impotence, ailment
noc	S	pl	F	R4e#	#	7	night
moc	S	pl	F	R4e#	#	17	power
pieśń	S	pl	F	R4y#	#	1	song
sieć	S	pl	F	R4y#	#	4	net
myśl	S	pl	F	R4y#	#	4	thought
wieść	S	pl	F	R4y#	#	8	news
rzecz	S	pl	F	R4y#	#	12	thing

**Tab. 3.3: Frequencies of i-stem feminine lemmas with different nom./acc. pl. suffixes in WMat.**

<sup>15</sup> Properly speaking, *R4y#* can only be considered the original *accusative* plural ending of the i-stems (< \*-i-ns), but already in Common Slavic times it was carried over to the nominative on the analogy of other feminines which had nom. pl.=acc. pl. (see Bielfeldt, 1961:137-138).

As Table 3.3 shows, WMat lemmas with either suffix show comparable frequencies. However, that R4e# has become much more widespread is clear from a comparison with the amount of types in GMat, where *noc* ‘night’ is the only i-stem noun to show R4e#:

Lemma	Pos	num	gend	suf	lemsuf	f(lemma) in G	English
<i>noc</i>	S	pl	F	R4e#	#	8	night
<i>majątność</i>	S	pl	F	R4y#	#	2	possession
<i>pieśń</i>	S	pl	F	R4y#	#	2	song
<i>sieć</i>	S	pl	F	R4y#	#	3	net
<i>twarz</i>	S	pl	F	R4y#	#	4	face
<i>myśl</i>	S	pl	F	R4y#	#	4	thought
<i>niemoc</i>	S	pl	F	R4y#	#	4	impotence, ailment
<i>wieść</i>	S	pl	F	R4y#	#	7	news
<i>ciemność</i>	S	pl	F	R4y#	#	8	darkness
<i>rzecz</i>	S	pl	F	R4y#	#	16	thing
<i>moc</i>	S	pl	F	R4y#	#	17	power

**Tab. 3.4: Frequencies of i-stem feminine lemmas with different nom./acc. pl. suffixes in GMat.**

This change has therefore been recognized automatically with ease, but the phenomenon’s token-frequencies are somewhat at odds with a traditional grammatical description. This is likely due to the nature of the Bible text and the textual medium in general: token frequency in written general language may well have a strong effect on the retention or replacement of irregular suffixes, but the influence of spoken language in this is far more crucial, especially for earlier periods with less literacy. Thus the word ‘power’ may be more frequent than ‘thing’ in the Gospel of Matthew, but this is unlikely to be true in general. However type counts do reveal the progressive takeover of the new suffix of more and more lemmas quite clearly, also independently of lemmas appearing in parallel, from 1/11 types in GMat to 5/10 in WMat.

### 3.3 Overview of Systematic Nominal Suffix Workloads

Another way of investigating inflectional morphology is to examine not which suffixes signify a grammatical category (e.g. changes in the suffixes expressing genitive or instrumental masculine), but rather, in the spirit of Jespersen’s Systematic Grammar (1924: 30-57), to ask what roles each suffix plays in the language. Again the discussion will be limited here to the suffixes of common nouns. Figure 3.5 gives the frequency of

each of the major nominal suffixes (very rare irregular suffixes, scoring less than 10 hits in both corpora, have not been considered), and how often they express which cases.

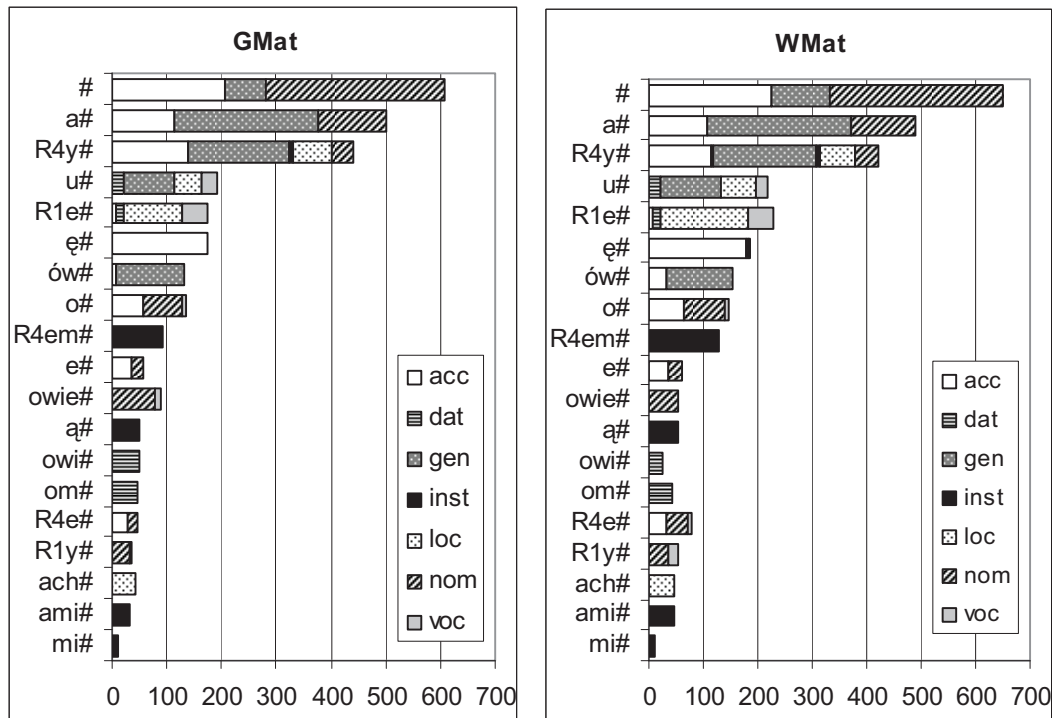


Fig. 3.5: Distribution of nominal suffixes and cases in both corpora.

The overall similarity of the distributions despite the limited size of the corpus stems from the fact that both texts share essentially the same content, but there are some subtle differences. For example, the development of the masculine personal plural discussed in section 3.2.1 can be seen in the bar for the suffix *ów#* (seventh bar from the top): the proportion of accusatives grows substantially here (the white part of the bar). The corresponding drop in the older, ousted *R4e#* accusatives, which one would expect, is partly obscured by the development of the previously discussed feminine i-stem plural in *R4e#* (section 3.2.5), but a small drop in *R4y#* accusatives (the soft allomorph of the ousted suffix) is actually present, from 137 to 114.

We can also notice that the suffixes marking the instrumental are all unambiguous (solid black bars) except for *R4y#*: the suffixes *a#*, *ami#*, *mi#* and *R4em#* mark only the instrumental, and the instrumental would be marked only by them, if not for the tiny black sliver in the *R4y#* bar (third from the top). This explains the pressure to



lose this ending as mentioned in section 3.2.3, which has made the instrumental an unambiguous category in the modern nominal declension.

The most ambiguous forms are the nominative, accusative and genitive in # (the ‘zero’ suffix) and a#. These are the most common grammatical cases, also typically marked by the shortest, least complex suffixes. This fits the long standing hypothesis that the lower informativity of common categories is correlated with the simplification and truncation of their phonological expression (for an analysis of Indo-European morphology along these lines see Martinet, 1962: 149-154). Most ambiguities involving these suffixes are resolved through nouns’ genders and number: a# serves for nom. in the feminine singular and nom. and acc. in the neuter plural, while it signifies genitive in the neuter singular. # signifies nominative for singular masculines, but genitive in the feminine and neuter plural. Thus the only possibly problematic ambiguity, given that speakers easily recognize the gender of a noun, is the neuter genitive singular versus nominative/accusative plural (the syncretism of nom. and acc. in all neuter forms is systemic, as in all Indo-European languages). This ambiguity can usually be resolved syntactically (e.g. through congruence, argument structure, etc.). If there were an alternative, somewhat frequent neuter form for either category we might still expect it to gain dominance, but such a form does not exist at present, leaving the ambiguity in place.

The decline and restriction of the masculine nominative/vocative plural in owie# and the corresponding rise of R1y# in these functions may correspond to actual historical phenomena (Rospond, 2003: 131-132), but are only noticeable in the graphs on account of the frequent lemmas ‘Pharisee’ and ‘Saducee’, which have lost the owie# suffix. This finding underscores the problem of content related lexical bias in a small corpus, and should therefore be interpreted carefully. Nonetheless, the lemma type counts, irrespective of the infrequent lemmas behind them, have also shown owie# to have become more restricted in WMat (12 lemmas in G vs. 6 in W, cf. section 3.2.4), and this process can be correctly identified in the end.

A point that was previously impossible to identify is the increase in the use of the locative singular in R1e# (from 107 instances to 160). Since this change does not involve the substitution of a suffix marking the same form it could not be identified by the suffix change query in section 3.2. The reasons for this distributional change seem to be varied,

but in any case relate to more use of the locative: one is a preference for the static locative in WMat over the dynamic accusative in GMat (e.g. G: *na wschód* ‘(we saw his star) towards the West’ versus W: *na Wschodzie* ‘in the West’, Mat. 2:2), another is the use of the temporal genitive in the old corpus and a prepositional locative in the new (G: *onejże godziny*, lit. ‘of that hour’ versus W: *w tej godzinie* ‘in that hour’, Mat. 8:13). This sort of finding, which can easily be missed by concentrating on outward changes in individual forms, is a good example of the kind of directions a corpus-based distributional examination can reveal for study. However at this stage, insight into the causes and meaning of such phenomena can only be gained by manually examining their occurrences. More advanced methods for detecting the nature of changes will be discussed in the following two chapters.

### 3.4 Summary and Evaluation

In this chapter I have reviewed the changes in nominal inflectional morphology as evidenced by differences in the suffix annotation of the parallel corpora, and discussed and compared them with traditional accounts of the development of Polish morphology in historical grammars. Section 3.2 concentrated on evidence pointed out by the query in Table 3.1, which was able to automatically identify the existence of historical processes in suffixal morphology by comparing minimal pairs of tokens, with identical annotations except for their suffix fields. Subsequent examinations of the corpus distributions of these variant suffixes were additionally able to quantitatively substantiate these changes, and give an idea of their relative importance within a grammatical category.

In section 3.3 I took advantage of the high comparability of case distributions in the corpora to create the graphic representations in Figure 3.5 and their analyses in terms of differences in nominal inflectional morphology. In several cases, the same facts as in section 3.2 could be revealed from a different angle, such as the appearance of the masculine personal plural genitive-accusative. The analysis of all suffixes in the context of other suffixes has the additional power to discover distributional phenomena that don’t produce pairs of parallel items with different forms, such as the increase in the use of the locative singular in R1e#. It also gives an overview dimension to the analysis which helps in identifying phenomena involving the functional load of morphological markers and

their possible motivation, as in the identification of the isolated status of the ambiguous instrumental plural suffix R4y#. The results overall show a system stabilizing after syncretism and the disuse of some inflectional classes have left substantial ambiguity. Surviving forms from older inflectional patterns are harnessed to create more systematic distinctions between the genders and cases, leading to less ambiguity, and to declensions incorporating the natural semantic categories of sex and animacy (masculine personal, animate and inanimate) on top of the old grammatical one of arbitrary gender.

Evaluating the accuracy of the results is in some respects difficult, since it is not clear what one expects to find in a study of morphological inflection. As far as precision is concerned, it seems clear that every one of the results in Table 3.1 corresponds to a real change in nominal morphology – if digitization and tagging errors can be ruled out, it is obvious that two suffixes for the same form are interesting, and demand an explanation. However, automatic queries cannot deliver an interpretation of these results. The most extreme case illustrating this problem is the accusative masculine personal suffix # for *maq̣z* ‘husband’ on line 9 of Table 3.1: this form was by no means regular or productive in either period of this corpus, and represents a frozen relict in the fixed expression *wyjść za maq̣z* ‘to marry (a man)’. This is however indistinguishable from other, productive changes that are observed, except in the low token and especially type frequency (the latter is = 1). Yet no clear cut criterion can be given for this distinction. Other findings range from pervasive regular changes (e.g. the MP accusative plural, section 3.2.1) to widespread but irregular (e.g. loss of *owie#* plural, section 3.2.4). Results are thus 100% ‘precise’ for the distinction ‘change : no change’ (perhaps with the exception of ‘to marry (a man)’), but not as fine-grained as one could wish.

As for recall, the question of what one expects to find is even more complex. In terms of inflectional tables, we find most of the changes in the major inflectional classes: in the o-stem masculines and neuters, development of the plural instrumental and MP accusative (hard and soft) are recognized, as well as the alternate nominative plural and genitive singular forms in the masculine. The feminine a-stems underwent no changes between the two stages, but the i-stem nominative/accusative plurals did, which was successfully detected. The only changes not detected concern some of the dual forms, which were already becoming obsolete in Middle Polish: the death of the dual

instrumental suffix in natural duals was only partly identified in row 6 of Table 3.1, where an old soft instrumental form *uszyma* ‘with (both) ears’ was found. The corresponding hard forms in oma# (e.g. *rękoma* ‘with (both) hands’) was only recognized in the examination of the distributions in Figure 3.4, where the hard oma# suffix appears only in the GMat graph, but not in WMat. The other two dual case forms are not found, since they appear only in natural duals even in Middle Polish, necessitating a sentence with ‘hands’, ‘eyes’ etc. in the appropriate case. As mentioned in section 3.2.3, the only such instances in the corpus have the nom./acc. pl. of the word for ‘hands’, which is *ręce*, which happens to be the only regularly preserved dual form in Modern Polish, and is thus not detected as a change. There is thus some indication of the death of the dual forms in Table 3.1 and Figure 3.4, but its limited extent is a direct result of the underlying corpus. Otherwise, it has been possible to detect all the changes one would expect to find in light of historical grammars.

While the trends in this corpus thus seem to be generally in line with traditional descriptions, it would have been desirable to compare changes in the distributions of suffixes across language stages with larger monolingual corpora. Because of the lack of a comparable electronic Middle Polish corpus, this has not been possible. Although the sublanguage of both Polimath corpora can undoubtedly be expected to differ somewhat from the general language quantitatively in many respects, future studies on small, parallel diachronic corpora could still greatly benefit from such a corpus for comparison, especially if it would provide normalized forms to facilitate comparisons. This direction holds the most exciting prospects, because it promises not only to try and replicate results from traditional studies based on empirical and reproducible data and methods, but also to enrich our knowledge further with quantitative accounts of diachronic change. At this point, however, quantitative results are limited to the scope of a case study, and it is only possible to evaluate the search for morphological change in this chapter in terms of a qualitative ‘binary’ detection of which suffixes underwent change. To this extent, the findings seem to provide good accuracy for the inventory of nominal morphology changes, but their significance can only be understood in the context of the actual data standing behind changed pairs, including token frequencies for suffixes, but also the identities and semantic extensions (e.g. *human male*) of lemma types they are attached to.

## **4 Changes in Verbal Lexis and Word Formation**

In this chapter, I will apply cooccurrence measures across parallel corpora in order to extract relationships between lexical items in both language stages. My goal will be to automatically create a concordance of the best WMat correspondence for each verbal lemma in GMat, and to identify and classify different categories of corresponding pairs on corpus-based criteria, with as little human intervention as possible. In this classification, I will be particularly interested in identifying correspondences that exhibit changes in verbal word formation. The envisioned end product of this investigation is a proportional case study representation of the lexical similarities and differences of verbal items between the two diachronically disparate parallel texts. As the results below will show, verbal lexis often remains constant across texts – most often completely, i.e. verb pairs have identical lemmas, but sometimes partly, e.g. using a common root in different lemmas, or especially often, having the same stem and differing only in prefixes.

### **4.1 Choosing Items**

As already mentioned in section 1.3, cooccurrence measures can be used to find items which appear in parallel aligned section consistently, forming likely candidates to be translations of each other in a translation corpus. In theory, items can be word forms, lemmas, or even morphological or syntactic features depending on the research question being asked. Because of the rich inflection in Polish, I will use lemma itemization, and not word forms. However, since often single tokens may be paralleled (or in a sense ‘translated’) by multiple tokens, it is necessary to decide how many tokens to consider in forming ‘items’. In order to do so, I will use the notion of collocations. Collocations are understood as sequences of multiple tokens whose semantic or syntactic properties cannot be predicted from their components (Evert, 2005: 17), and which crucially in this context, may therefore have their own corresponding translations independently of their components. To identify collocations in each corpus I will use the *z*-score (Berry-Rogghe, 1973), a well established measure which has the advantage of applying to both contiguous and non-contiguous token sequences (for an evaluation of the *z*-score against other measures see Pearce, 2002). The measure is calculated as follows (Oakes, 1998:

163-166): Given two items  $a$  and  $b$ , which appear  $A$  and  $B$  times respectively in a corpus of  $N$  items, including  $C$  times in which they appear in the span of  $S$  items from each other (i.e. in collocation), the probability of  $b$  appearing where  $a$  does not appear in the span is  $p$  (on the left in Figure 4.1), and the expected number of cooccurrences is given by  $E(c)$  (in the middle). The  $z$ -score (on the right) determines whether the discrepancy between  $C$  and its expected value is statistically significant ( $z > 2.576$  for a 1% significance level):

$$p = \frac{B}{N - A} \quad E(c) = p \cdot A \cdot S \quad z = \frac{C - E(c)}{\sqrt{E(c)(1 - p)}}$$

**Fig. 4.1: Formulas for  $p$ ,  $E(c)$  and the  $z$ -score.**

I compute this measure using SQL, by joining one corpus table onto itself using consecutive token identifiers within a span of 2 items (an arbitrarily chosen span with easily manageable size) as the join property. This produces a table of all possible pairs of items appearing within a span  $S=2$ . I then use a query on the resulting table to count distinct pairs, giving  $C$ , and compute  $z$  in a further query using type counts from the corpus for  $A$  and  $B$ . This process is repeated for both corpora, and results appearing less than twice or containing punctuation are heuristically filtered out. The result is a  $z$ -score for each two collocation candidates appearing in the corpus within  $S=2$  (see Table 4.1).

<b>a</b>	<b>b</b>	<b>A</b>	<b>B</b>	<b>C</b>	<b>p</b>	<b>E(C)</b>	<b>z</b>
<i>przedni</i>	<i>kapłan</i>	18	28	18	0.001181434	0.0425316455696	87.1255863013012
<i>syn</i>	<i>na</i>	101	229	3	0.009696405	1.9586738366431	0.747689946817026

**Tab. 4.1:  $z$ -score of collocation candidates in GMat.**

The first entry in the table, with the lemmas *przedni* ‘front’ and *kapłan* ‘priest’, receives a high  $z$ -score, since the appearance of *przedni* conditions the following *kapłan* in this corpus ( $A=C=18$ ). This is because they often appear together in the phrase *przedniejszy kapłan* ‘foremost priest’. The coincidental sequence in the second entry, *syn* ‘son’ and *na* ‘on’, receives a low score. Note that this collocation extraction method is quite crude, putting recall before precision; many collocations are a direct result of the small corpus size and its specific language (e.g. the fact that ‘foremost priest’ is a significant collocation, which may not hold for the general language).

## 4.2 Identifying Parallels

Once the list of collocations is established, the correlations between items in parallel sections can be tested across the corpora. For this purpose I will use Daille’s (1995: 36-37) Cubic Association Ratio (sometimes called mutual information cubed, or MI3 for short) which seems to perform well, though another measure tested, Log Likelihood (Dunning, 1993), has shown very similar, though subjectively slightly worse results. MI3 gives a score between plus and minus infinity of how likely we are to find item *b* in a parallel section given that item *a* appears in the source section. To compute it, all items that appear in parallel aligned sections must be paired in all possible configurations, i.e. the first item in the first verse is paired with each item in the parallel version of the first verse, then likewise for the second item in the first verse, and so on for all verses. Note that for each item, parallels are drawn only from the same verse. Given the total amount of pairs  $f(a)$  in which *a* appears, the amount of pairs  $f(b)$  in which *b* appear, and  $f(a\&b)$ , the amount of pairs where both appear, MI3 is given by  $\log_2 (f^3(a\&b) \cdot N / f(a) \cdot f(b))$  where N is the total amount of pairs (see also Oakes, 1998: 170-172). Computing MI3 directly from the corpora can again be done using SQL – the two tables containing both corpora are joined on parallel verse identifiers, producing the desired pairing of all parallel items with the same verse identifier. Punctuation and other token-types with a frequency of more than 1% in either corpus (e.g. “function words” like ‘and’ etc.) are of no direct interest for this lexical study, and their entries are therefore eliminated.

This procedure results in a table listing the association strength between each two lemmas or collocations that appear in parallel aligned sections (see Table 4.2). Because of the similarity of the texts in the corpus, as well as the normalized orthography, matching items may often be identical (row 1). Also, collocations may match with single lemmas (row 2): the lemmas *przedni* ‘front’ and *kaplan* ‘priest’ from the collocation example in section 4.1 are paralleled in the new corpus by the single lemma *arcykaplan* ‘archpriest’.

Row	a (GMat)	b (WMat)	Sense	A	B	C	MI3
1	<i>słowo</i>	<i>słowo</i>	word	343	585	24	14.001
2	<i>przedni kaplan</i>	<i>arcykaplan</i>	chief priest	441	237	19	13.931

**Tab. 4.2: Matching lemmas and collocations between corpora.**



### 4.3 Changes in Verbs and Verb Substitution Types

The following sections deal with an examination and classification of correspondences between verbal lemmas across the corpora as they appear in the concordance illustrated in Table 4.2. The correspondence classes will be described first, followed by the methods for their acquisition, while the relative proportions of the different resulting classes will be shown and discussed in the summary in section 4.4. As a most basic classification, I will distinguish between verbal lemmas that are paralleled by other verbal lemmas and verbal lemmas that are paralleled by non-verbal lemmas. In the latter case, we may find that the best MI3 correspondence of a verbal lemma is for example a noun, adjective or collocation. If a verbal lemma is paralleled by another verbal lemma, the two lemmas may be identical, forming the class of unchanged verbal lemmas. If there is a non-identical verb-verb correspondence, we may find that a verb has been substituted by a completely unrelated verb. In many cases, however, only a part of the verb is substituted, in which case we can check which parts of the lemma have been replaced, such as prefixes but not the stem or the stem but not prefixes, or even parts of the stem, such as changes in suffixes or vowel gradation, which can produce a new lemma from the same root (i.e. the abstract lexical morpheme from which verb stems are formed). As we shall see in section 4.3.2, even the rules for combining these elements may change. Figure 4.2 illustrates different types of verb-verb partial correspondences with some basic examples.

	<b>GMat</b>		<b>WMat</b>	
<b>Prefix change:</b>	<i>na-śmiać</i>	:	<i>wy-śmiać</i>	‘ridicule’
	at-laugh		out-laugh	
<b>Stem change:</b>	<i>wy-gnać</i>	:	<i>wy-pędzić</i>	‘drive out’
	out-chase		out-rush	
<b>Suffix change:</b>	<i>za-bieżeć</i>	:	<i>za-biec</i>	‘run across’ (both from root <i>bieg</i> )
	beyond-run		beyond-run	

Fig. 4.2: Examples of corresponding verb pairs with different parts substituted.

By classifying verb correspondences I will attempt to give a picture of how much has changed between the two texts in terms of verbal lexis, and what sorts of change are most prevalent. Partial changes are of particular interest in that they affect a larger portion of the lexicon by renegotiating the linguistic value of the prefixes and stems in question,



which may be used in many different lemmas, and at the same time influencing all the other prefixes and stems in opposition to them (particularly the other elements involved in the substitution in that instance).

In order to find what has changed from the diachronic perspective of the verbs in the older text, I will filter the item concordance produced in section 4.2 to show the best MI3 correspondence of all verbal lemmas in GMat. If multiple parallels have identical MI3 scores but one is a pair of identical lemmas, I assume this is the best parallel (the null hypothesis is that no change has occurred). This simplification ignores polysemy, or more accurately in this paradigm, the case where one verb may have multiple parallels in different contexts. I will only be considering the single most consistent match for each lemma, the reason being that the limited context offered by this small corpus is insufficient to distinguish senses on corpus data alone. A sense-annotated corpus could in theory be used to regard different senses of verbs as distinct lemmas (for further discussion see section 4.4).

It is important to make the fact explicit that changes such as the substitution of prefixes etc. may also modify the meaning of a verb to such an extent that a pair found to correspond is only parallel in a particular use or sense. In treating a pair as parallel I assume a measure of semantic uniformity between the texts by virtue of their forming a parallel corpus: the statistically soundest matching pair only means that one item was most consistently chosen over the other in the context of this corpus, be it for reasons of a total ousting of the old form through language change or merely stylistic variation between competing items (cf. section 1.3).

In order to help categorize the retrieved pairs I will use some additional information and a number of functions. Firstly, I will use part-of-speech information in order to distinguish between verb-verb and other correspondences. Secondly, I will compute Levenshtein Distance (LD, Levenshtein, 1966) between each two items. This string similarity measure checks how many character insert, delete or replace operations are required to transform one string into another. For example, in order to transform the item *naśmiać* in Figure 4.2 above into its parallel *wyśmiać*, the first two characters must be replaced, resulting in LD=2. Items with zero LD are identical, and represent non-change of a lemma. High LD is characteristic of total replacement of a lemma, probably

including the root, while low values signify a partial change. I will also check where the difference between lemmas is: at the left of the strings (possible prefix change), at the right (stem change), or both. I define two sets of functions for this purpose: *LeftChangeIndex* and *RightChangeIndex*, which return for each lemma the amount of characters left in a string once the first difference has been detected, starting from the left and from the right respectively; and *LeftIdentIndex* and *RightIdentIndex*, which return the amount of identical characters on either end of the strings. Table 4.3 illustrates some pairs and the values of these functions.

a (GMat)	b (WMat)	Sense	MI3	LD	LId	RIId	aLC	aRC	bLC	bRC
<i>obwarować</i>	<i>zabezpieczyć</i>	guard	10.357	10	0	1	9	8	12	11
<i>pełnić</i>	<i>spełniać</i>	fulfil	10.575	2	0	1	6	5	8	7
<i>pogrześć</i>	<i>pogrzebać</i>	bury	12.365	2	6	1	2	7	3	8
<i>zadziwić</i>	<i>zdziwić</i>	amaze	11.302	1	1	6	7	2	6	1
<i>wziąć</i>	<i>wziąć</i>	take	15.513	0	5	5	0	0	0	0

**Tab. 4.3: Examples of verb change types with string comparison measures. High LD indicates total replacement, LD=0 means the lemma was retained. Low LD signals partial change, in prefix and suffix, suffix only or prefix only (middle 3 rows).**

### 4.3.1 Retained Verbal Lemmas

The simplest class of verbs to identify is the group of unchanged verbs. To find these the distinct GMat lemmas are counted whose best match is the same lemma in WMat. This can be done by specifying that both lemma fields must be identical (a=b), or by looking for zero Levenshtein distance. There are a total of 354 such verbs, which may be found in Table 4.4 across the next pages (in order to save space, only the single lemma and the MI3 rating with which it matches the same lemma are given).

Lemma	MI3	Lemma	MI3	Lemma	MI3	Lemma	MI3	Lemma	MI3
<i>być</i>	23.6	<i>wydać</i>	13.8	<i>zadusić</i>	12.2	<i>sprzedawać</i>	11.1	<i>pamiętać</i>	10.2
<i>przyjmować</i>	17.7	<i>wyganiać</i>	13.7	<i>pocieszyć</i>	12.2	<i>siąść</i>	11.1	<i>wzywać</i>	10.2
<i>mieć</i>	17.4	<i>usidlić</i>	13.7	<i>rozpuścić</i>	12.2	<i>nasycić</i>	11.1	<i>zdobić</i>	10.2
<i>przysiąć</i>	17.2	<i>rozsyptywać</i>	13.7	<i>prześadować</i>	12.1	<i>wjechać</i>	11.1	<i>odłączyć</i>	10.1
<i>mówić</i>	16.9	<i>pozdrawiać</i>	13.7	<i>oskarżyć</i>	12.1	<i>dołamać</i>	11.1	<i>zdumiewać</i>	10.1
<i>przysięgać</i>	16.6	<i>ukrzyżować</i>	13.7	<i>przybliżyć</i>	12.1	<i>palić</i>	11.1	<i>uciec</i>	10.1
<i>pić</i>	16.6	<i>bić</i>	13.6	<i>poznać</i>	12.1	<i>zagasić</i>	11.1	<i>wysłuchać</i>	10.1
<i>przyjść</i>	16.4	<i>cudzołożyć</i>	13.6	<i>wytrwać</i>	12.1	<i>ukazać</i>	11.1	<i>spalić</i>	10.0
<i>zgorszyć</i>	16.1	<i>szemrać</i>	13.5	<i>uwolnić</i>	12.1	<i>radować</i>	11.0	<i>zgromadzić</i>	9.9
<i>milować</i>	16.1	<i>skłonić</i>	13.5	<i>umrzeć</i>	12.1	<i>chodzić</i>	11.0	<i>obcinać</i>	9.9
<i>usłuchać</i>	16.0	<i>przyodziać</i>	13.5	<i>dawać</i>	12.1	<i>wrzucić</i>	11.0	<i>przechodzić</i>	9.9

<i>móc</i>	16.0	<i>zabłąkać</i>	13.5	<i>siedzieć</i>	12.0	<i>gromić</i>	11.0	<i>potępić</i>	9.9
<i>uczynić</i>	16.0	<i>spać</i>	13.5	<i>upodobać</i>	12.0	<i>posiać</i>	11.0	<i>zaduszać</i>	9.9
<i>czynić</i>	16.0	<i>zgrzytać</i>	13.3	<i>opętać</i>	12.0	<i>postawić</i>	11.0	<i>najmować</i>	9.8
<i>chcieć</i>	15.9	<i>znać</i>	13.3	<i>leżeć</i>	12.0	<i>obchodzić</i>	11.0	<i>odwizać</i>	9.8
<i>prosić</i>	15.9	<i>posłać</i>	13.2	<i>bluźnić</i>	12.0	<i>przejrzeć</i>	11.0	<i>uwiązać</i>	9.8
<i>służyć</i>	15.9	<i>uzdrawiać</i>	13.2	<i>przenieść</i>	12.0	<i>wyłączyć</i>	11.0	<i>żałować</i>	9.8
<i>uwierzyć</i>	15.9	<i>wychodzić</i>	13.2	<i>złorzeczyć</i>	11.9	<i>mniewać</i>	10.9	<i>przyodziewać</i>	9.7
<i>dać</i>	15.8	<i>dostąpić</i>	13.2	<i>wyłożyć</i>	11.9	<i>odwracać</i>	10.9	<i>zaćmić</i>	9.7
<i>widzieć</i>	15.6	<i>trwożyć</i>	13.2	<i>policzkować</i>	11.9	<i>pożyczyć</i>	10.9	<i>uchwycić</i>	9.6
<i>odpowiadać</i>	15.6	<i>uderzyć</i>	13.2	<i>ślać</i>	11.9	<i>święcić</i>	10.9	<i>odwalić</i>	9.6
<i>wziąć</i>	15.5	<i>milczeć</i>	13.2	<i>ufać</i>	11.9	<i>rozejść</i>	10.9	<i>pojednać</i>	9.6
<i>odpuścić</i>	15.5	<i>zgrzeszyć</i>	13.1	<i>znajdować</i>	11.9	<i>nająć</i>	10.8	<i>założyć</i>	9.6
<i>słyszeć</i>	15.4	<i>oddać</i>	13.1	<i>musieć</i>	11.9	<i>stworzyć</i>	10.8	<i>podnieść</i>	9.6
<i>modlić</i>	15.3	<i>znosić</i>	13.1	<i>wpaść</i>	11.9	<i>ubiczować</i>	10.8	<i>minąć</i>	9.6
<i>jeść</i>	15.3	<i>prorokować</i>	13.0	<i>przywieść</i>	11.8	<i>obrócić</i>	10.8	<i>napęlić</i>	9.6
<i>poniżać</i>	15.3	<i>podeptać</i>	13.0	<i>dokończyć</i>	11.7	<i>odjechać</i>	10.8	<i>ochrzcić</i>	9.5
<i>wywyższać</i>	15.3	<i>łaknąć</i>	13.0	<i>kupić</i>	11.7	<i>oszacować</i>	10.8	<i>rozwalać</i>	9.5
<i>zaprzec</i>	15.2	<i>odebrać</i>	13.0	<i>zatrwożyć</i>	11.7	<i>porywać</i>	10.8	<i>biczować</i>	9.5
<i>iść</i>	15.2	<i>napisać</i>	13.0	<i>kazać</i>	11.7	<i>przemienić</i>	10.8	<i>ogrodzić</i>	9.5
<i>zwiesić</i>	15.0	<i>pozdrowić</i>	13.0	<i>odchodzić</i>	11.7	<i>wykopać</i>	10.8	<i>wkopać</i>	9.5
<i>potępiać</i>	14.9	<i>ściąć</i>	13.0	<i>rozpraszać</i>	11.7	<i>żeńić</i>	10.8	<i>przejść</i>	9.5
<i>rozwiązać</i>	14.9	<i>przykazać</i>	12.9	<i>zapiać</i>	11.7	<i>umieć</i>	10.8	<i>przełożyć</i>	9.5
<i>zabić</i>	14.8	<i>wołać</i>	12.9	<i>pojmać</i>	11.7	<i>zakryć</i>	10.8	<i>brać</i>	9.5
<i>powiadać</i>	14.8	<i>gardzić</i>	12.9	<i>rozumieć</i>	11.7	<i>tonąć</i>	10.7	<i>przyłożyć</i>	9.4
<i>wypełnić</i>	14.8	<i>zginąć</i>	12.9	<i>wierzyć</i>	11.7	<i>złąć</i>	10.7	<i>wielbić</i>	9.4
<i>nieść</i>	14.8	<i>kraść</i>	12.9	<i>obfitować</i>	11.6	<i>rwać</i>	10.7	<i>wysławiać</i>	9.4
<i>wracać</i>	14.8	<i>ukraść</i>	12.8	<i>zgromić</i>	11.6	<i>stawiać</i>	10.7	<i>popęlić</i>	9.4
<i>opuścić</i>	14.8	<i>zwyknąć</i>	12.8	<i>zbawić</i>	11.6	<i>wykładać</i>	10.7	<i>wschodzić</i>	9.4
<i>usłyszeć</i>	14.8	<i>znaleźć</i>	12.8	<i>kłaść</i>	11.6	<i>zapalać</i>	10.7	<i>odłączyć</i>	9.4
<i>powiedzieć</i>	14.8	<i>szukać</i>	12.7	<i>przestępować</i>	11.6	<i>pokazać</i>	10.6	<i>ujść</i>	9.4
<i>błogosławić</i>	14.7	<i>uragać</i>	12.7	<i>uschnąć</i>	11.5	<i>przepowiedzieć</i>	10.6	<i>upadać</i>	9.3
<i>przymuszać</i>	14.7	<i>zasmucić</i>	12.7	<i>skosztować</i>	11.5	<i>krzyknąć</i>	10.6	<i>upleść</i>	9.3
<i>wyjąć</i>	14.7	<i>pocałować</i>	12.7	<i>zmieszać</i>	11.5	<i>dostawać</i>	10.6	<i>nadstawić</i>	9.3
<i>przystąpić</i>	14.7	<i>zabijać</i>	12.7	<i>łamać</i>	11.5	<i>mierzyć</i>	10.6	<i>plakać</i>	9.3
<i>wiedzieć</i>	14.7	<i>chrzcić</i>	12.7	<i>blądzić</i>	11.5	<i>odjeżdżać</i>	10.6	<i>przemóc</i>	9.3
<i>dotknąć</i>	14.6	<i>odpowiedzieć</i>	12.6	<i>czerwienić</i>	11.5	<i>odmierzyć</i>	10.6	<i>rozkazywać</i>	9.3
<i>stać</i>	14.6	<i>pokutować</i>	12.6	<i>przyłecić</i>	11.5	<i>przypatrywać</i>	10.6	<i>utopić</i>	9.2
<i>przeminąć</i>	14.5	<i>przeprowadzić</i>	12.6	<i>rozkazać</i>	11.5	<i>wylać</i>	10.5	<i>zawiesić</i>	9.2
<i>upaść</i>	14.5	<i>wspomnieć</i>	12.6	<i>przestrzegać</i>	11.5	<i>obciążyć</i>	10.5	<i>pozyskać</i>	9.2
<i>wypuścić</i>	14.4	<i>karmić</i>	12.5	<i>świadczyc</i>	11.4	<i>gasnąć</i>	10.5	<i>przebywać</i>	9.2
<i>wyciągnąć</i>	14.4	<i>przychodzić</i>	12.5	<i>pożądać</i>	11.4	<i>prząść</i>	10.5	<i>zabraniać</i>	9.1
<i>związać</i>	14.4	<i>strzec</i>	12.5	<i>czcić</i>	11.4	<i>rosnąć</i>	10.5	<i>odpoczywać</i>	9.1
<i>słuchać</i>	14.4	<i>nawrócić</i>	12.5	<i>zostawić</i>	11.4	<i>rozłączać</i>	10.5	<i>rodzić</i>	9.1
<i>wyznać</i>	14.2	<i>oziębnać</i>	12.5	<i>pluć</i>	11.4	<i>złączyć</i>	10.5	<i>pożerać</i>	9.1
<i>skrócić</i>	14.2	<i>połykać</i>	12.5	<i>ująć</i>	11.4	<i>chwalić</i>	10.5	<i>pobłogosławić</i>	9.0
<i>pościć</i>	14.1	<i>powiesić</i>	12.5	<i>uniżyć</i>	11.4	<i>gorszyć</i>	10.4	<i>pobiełać</i>	8.9
<i>oczyścić</i>	14.1	<i>przecedzać</i>	12.5	<i>wypuszczać</i>	11.4	<i>odpuszczać</i>	10.4	<i>przylatywać</i>	8.9
<i>wchodzić</i>	14.1	<i>rozmnożyć</i>	12.5	<i>troszczyć</i>	11.4	<i>podziękować</i>	10.4	<i>solić</i>	8.9
<i>odejść</i>	14.1	<i>zebrać</i>	12.5	<i>zbudować</i>	11.3	<i>wyrzucić</i>	10.4	<i>związać</i>	8.9
<i>bać</i>	14.1	<i>zwać</i>	12.4	<i>włożyć</i>	11.3	<i>sposztrzec</i>	10.4	<i>pomagać</i>	8.8

<i>czytać</i>	14.1	<i>weszać</i>	12.4	<i>przynieść</i>	11.3	<i>zagarniać</i>	10.4	<i>wzbudzić</i>	8.8
<i>uzdrowić</i>	14.1	<i>zbierać</i>	12.4	<i>kupować</i>	11.3	<i>zapaść</i>	10.4	<i>obejść</i>	8.8
<i>kusić</i>	14.0	<i>zmartwychwstać</i>	12.3	<i>sprzeciwiać</i>	11.3	<i>świecić</i>	10.3	<i>ożyć</i>	8.7
<i>ujrzyć</i>	14.0	<i>potrzebować</i>	12.3	<i>wątpić</i>	11.2	<i>umyć</i>	10.3	<i>skonać</i>	8.7
<i>przyjąć</i>	13.9	<i>kolatać</i>	12.3	<i>zaprosić</i>	11.2	<i>dręczyć</i>	10.3	<i>królować</i>	8.5
<i>otworzyć</i>	13.9	<i>policzyć</i>	12.3	<i>grać</i>	11.2	<i>żąć</i>	10.3	<i>wystąpić</i>	8.5
<i>pójść</i>	13.9	<i>nazwać</i>	12.3	<i>śpiewać</i>	11.2	<i>weseleć</i>	10.3	<i>nakarmić</i>	8.4
<i>sądzić</i>	13.9	<i>napoić</i>	12.3	<i>pochodzić</i>	11.2	<i>wiązać</i>	10.2	<i>odbierać</i>	8.4
<i>ratować</i>	13.9	<i>mleć</i>	12.2	<i>powstać</i>	11.2	<i>objawić</i>	10.2	<i>gniewać</i>	7.6
<i>wstać</i>	13.8	<i>rozślawić</i>	12.2	<i>wybrać</i>	11.2	<i>posyłać</i>	10.2		

Tab. 4.4: Query results for verbal lemmas best matching the same lemma.

The large number of verbs matching themselves is an indication of the relative accuracy of the performance of the correlation measure MI3 in this corpus, since identical matches are essentially certain to be correct parallels. The meaning of the large number of identical pairs will be discussed in section 4.4. In the following sub-sections, I will classify the remaining verbs, which have non-identical pairs.

### 4.3.2 Prefix Changes

Pairs exhibiting prefix change will have strings that are identical, except for some part of the string on the left side. A single Polish prefix can be between 1 character (*o-* in *o-budzić* ‘to rouse’) and 4 characters long (*przy-* in *przy-wołać* ‘to call, summon’), and although multiple prefixes can reach a combined length longer than 4 (e.g. *roz-* + *po-* in *roz-po-godzić* ‘clear up’), it will be assumed for the moment that prefixes occupy no more than these first 4 characters. Beyond the prefixes, the verb must have a stem left over which is at least one syllable long, with the minimal structure CVC, where the last C is occupied by the infinitive suffix <ć> (realized as <c> for stems ending with an underlying velar consonant). This means each lemma pair in this class must end with at least 3 identical characters<sup>16</sup>. The criteria for a prefix change are therefore set at *RightChangeIndex* < 5 in both lemmas (the space occupied by the prefixes), and *LeftIdentIndex* > 2, which is required in order to show the identical verb stem. Table 4.5 shows the result of a query for prefix substitutions using these criteria. The suggested different prefixes (marked in bold) are extracted automatically by taking the first *RightChangeIndex* number of characters on the left of the respective lemma field.

<sup>16</sup> This also applies to the stem of *iść/-jść* ‘go’, which exceptionally violates the minimal CVC structure.

The 58 results can be divided into 3 groups. The last six rows represent errors which stem from similar looking verbs not actually exhibiting a prefix change. One pair *na-lamać* ‘crack’ : *do-lamać* ‘break’ does differ in prefix only, but the match is incorrect: though similar in meaning and appearing together in the text, the correct match as far as the parallel text is concerned is found with an equal score further up the list: *na-lamać* : *nad-lamać*. The correct pair *smęcić* : *smucić*, both ‘to mourn’, actually represents the same word etymologically, but the latter is a Czech loan, with a typical /u/ for the nasal /ę/ (cf. Siatkowski, 1970: 13); nonetheless, *smę-* and *smu-* are not prefixes, and no prefix change is involved. *zmiłować* : *zlitować* ‘pity’ is also a correct pair with coincidentally similar endings and no prefix change. The remaining errors are match errors.

The other 52 verb pairs truly differ only in prefixes (including the borderline case *d-ufać* : *za-ufać* ‘to believe, trust’, where the verbs are related, and the new lemma has added a prefix, but the old lemma’s initial *d-* is not a transparent prefix), producing a

G lemma	W lemma	Sense	MI3	G lemma	W lemma	Sense	MI3
<i>wynijść</i>	<i>wyjść</i>	go out	15.24	<i>dufać</i>	<i>zaufać</i>	believe	10.35
<i>wnijść</i>	<i>wejść</i>	go in	14.35	<i>urosnąć</i>	<i>podrosnąć</i>	grow	10.00
<i>paść</i>	<i>upaść</i>	fall	13.57	<i>zwołać</i>	<i>przywzwać</i>	convene	9.90
<i>począć</i>	<i>zacząć</i>	begin	13.22	<i>wejrzyć</i>	<i>spojrzeć</i>	glance	9.89
<i>skryć</i>	<i>ukryć</i>	hide	12.80	<i>wyrozumieć</i>	<i>rozumieć</i>	understand	9.81
<i>stawić</i>	<i>wystawić</i>	stand (vt.)	12.70	<i>odnieść</i>	<i>zanieść</i>	carry	9.68
<i>uwinąć</i>	<i>owinąć</i>	wrap	12.69	<i>otrząsnąć</i>	<i>strząsnąć</i>	shake off	9.61
<i>wzrosnąć</i>	<i>wyrosnąć</i>	grow	12.25	<i>zawzwać</i>	<i>przywzwać</i>	call	9.53
<i>obudzić</i>	<i>zbudzić</i>	wake up	12.11	<i>rozszzerzać</i>	<i>poszerzać</i>	widen	9.37
<i>przydać</i>	<i>dodać</i>	add	12.08	<i>zgotować</i>	<i>przygotować</i>	prepare	9.08
<i>zaśpiewać</i>	<i>odśpiewać</i>	sing	11.88	<i>zamyślać</i>	<i>rozmyślać</i>	ponder	8.69
<i>poświęcać</i>	<i>uświęcać</i>	consecrate	11.76	<i>przeklinać</i>	<i>zaklinać</i>	curse	8.31
<i>poprzedzić</i>	<i>wyprzedzić</i>	precede	11.73	<i>pokalać</i>	<i>kalać</i>	defile	15.77
<i>naśmiać</i>	<i>wyśmiać</i>	ridicule	11.69	<i>żądać</i>	<i>zażądać</i>	desire	13.98
<i>nagotować</i>	<i>przygotować</i>	prepare	11.62	<i>zapięczętować</i>	<i>pieczętować</i>	seal	12.28
<i>osławić</i>	<i>zniesławić</i>	dishonor	11.52	<i>drżeć</i>	<i>zadrzeć</i>	tremble	12.25
<i>narodzić</i>	<i>urodzić</i>	be born	11.34	<i>maczać</i>	<i>umaczać</i>	wet	11.88
<i>zadziwić</i>	<i>zdziwić</i>	amaze	11.30	<i>rozumieć</i>	<i>rozumieć</i>	understand	11.71
<i>padać</i>	<i>spadać</i>	fall	11.25	<i>wiać</i>	<i>powiać</i>	blow	11.20
<i>nałamać</i>	<i>nadłamać</i>	crack	11.08	<i>podobać</i>	<i>spodobać</i>	please	10.70
<i>ubić</i>	<i>zbić</i>	beat up	11.08	<i>trząść</i>	<i>zatrząść</i>	shake	10.19
<i>strudzić</i>	<i>utrudzić</i>	tire	10.95	<i>mieszkać</i>	<i>zamieszkać</i>	dwel	9.59
<i>spytać</i>	<i>zapytać</i>	ask	10.93	<i>pytać</i>	<i>zapytać</i>	ask	9.59
<i>usiąść</i>	<i>zasiąść</i>	sit down	10.85	<i>zmiłować</i>	<i>zlitować</i>	pity	13.01
<i>naśmiewać</i>	<i>wyśmiewać</i>	ridicule	10.82	<i>smęcić</i>	<i>smucić</i>	mourn	12.36
<i>okrywać</i>	<i>przykrywać</i>	cover	10.58	<i>pragnąć</i>	<i>łaknąć</i>	desire/hunger	12.13
<i>przylączyć</i>	<i>połączyć</i>	join	10.46	<i>nałamać</i>	<i>dołamać</i>	crack/break	11.08
<i>umieść</i>	<i>wymieść</i>	sweep	10.37	<i>nasadzić</i>	<i>ogrodzić</i>	plant/fence	9.52
<i>wsiać</i>	<i>zasiać</i>	sow	10.37	<i>szpecić</i>	<i>pościć</i>	deface/fast	8.97

Tab. 4.5: Query results for verbal prefix changes.

precision of  $52/58 \approx 90\%$  correct prefix change pairs in the result set. As for recall, I assume that missing prefix changes must stem from one of two reasons. The first possibility is that a prefix change within the MI3 concordance was not identified by the string comparison criteria. A manual examination of all verb-verb correspondences has revealed only one such case, the pair *po-prze-wracać* : *po-wy-wracać* ‘to overturn’, which is due to a string-internal change of the second prefix being missed, violating the maximum length constraint of 4 characters. The other option is that a translation pair is missing in the correspondence table (because a better match could be found instead, etc.). In this case we may say the match was not well attested in the corpus under the current criteria in the first place, and can therefore be safely left out. This subsumes that correctly parallel MI3 pairs are taken to be the gold standard of what constitutes a match, and that only one, best parallel is allowed for each lemma; these assumptions will be revisited in the evaluation of the performance of MI3 in section 4.4. Thus, within this paradigm at least, recall is  $52/53 \approx 98\%$ , for an F-score of:  $F = 2 \cdot Pr \cdot Rc / (Pr + Rc) \approx 94\%$  for the classification criteria.

In this way, the parallel corpus can automatically deliver a fairly reliable list of parallel verbs differing only in prefixes, and the differing prefix strings. However, the second group of 11 verbs (marked gray), which exhibit an alternation between having some prefix and no prefix, can all be ascribed to grammatical, and not lexical differences – the prefixed verb is the perfective counterpart of the unprefixed verb. In these cases the new text uses a construction with a different grammatical aspect of the same verb, which entails substituting the lemma for one with the appropriate aspect. This incidentally reveals that the perfective form is probably showing the ‘default’ perfectivizing prefix<sup>17</sup>, with minimal semantic influence, which can be of lexicographic interest in itself. These pairs can be omitted from the query by specifying that the aspect of both lemmas must match (this is possible since the corpus is tagged for aspect). The remaining 41 pairs exhibit several types of interesting historical phenomena in the variation of verbal prefixation:

---

<sup>17</sup> As in other Slavic languages, prefixation is not only a means of forming perfective verbs, but can also change the meaning of the verb either slightly or substantially. For each verb, (at least) one prefix, which is said to change the meaning of the verb least, is considered the default perfective prefix. This prefix is unpredictable, and therefore given in dictionaries alongside a verb’s entry. For more on default prefixes and actual pair types see e.g. Włodarczyk and Włodarczyk (2001) and Swan (2002: 277, 281-285).



1. Use of prefixed perfective verbs instead of unprefixed, inherently perfective ones: *paść* : *u-paść* ‘fall’, *stawić* : *wy-stawić* ‘stand s.t. out, deploy’. Inherently perfective verbs are a semantically motivated morphological anomaly – almost all simplex (i.e. unprefixed) verbs are imperfective, but a few are perfective by virtue of their meaning (to fall or make something stand is understandably usually punctual). Adding a prefix has harmonized morphology and semantics.
2. Use of prefixed verbs with specialized senses vs. more general or polysemous verbs: *stawić* : *wy-stawić* ‘stand s.t. out, deploy’ (*stawić* has more senses outside this context), and conversely *wy-rozumieć* : *z-rozumieć* ‘understand’ (*wy-rozumieć* has a more specific sense of ‘fully understanding’). In the former case, a new verb selects a semantic subset of the meaning of the old verb, and in the latter, a more general verb is used since the older verb has gone out of use. In either case, the semantic narrowing effect of a prefix can be observed.
3. Different choices of default perfectivizing prefixes, which are still in competition today, e.g. *o-budzić* : *z-budzić* ‘rouse’, *s-pytać* : *za-pytać* ‘ask’, etc. The choice of default prefix is not completely fixed for all verbs in Modern Polish, and here the diachronic factor may not be the pertinent one.
4. Use of *wy-* ‘out’ to focus on resultativity of perfective verbs: *u-mieść* : *wy-mieść* ‘to sweep’, where the old *u-* prefix expresses an amount (‘how much swept’) and new *wy-* expresses direction instead (‘sweeping something out’). Also *na-śmiać* : *wy-śmiać* ‘to ridicule’ (‘laugh someone out’ instead of ‘laugh at someone’), and *po-przedzić* : *wy-przedzić* ‘to outpace, precede’, with a default *po-* replaced by *wy-*, perhaps expressing the preceded person is left out or behind (cf. Eng. *outrun*).
5. Change in morphotactics in prefixation of verbs with initial vowel: *wyn-ijść* : *wy-jść* ‘go out’, *wn-ijść* : *we-jść* ‘go in’. The latter two pairs historically have the same stem meaning ‘to go’ and the same prefixes *wy-* ‘out’ and *w-* ‘in’<sup>18</sup> in both corpora, but the older forms have an /n/ between prefix and stem. This is due to an Old Polish phonotactic rule inserting /n/ between prefixes and stems beginning with a vowel, which was generalized from two common prefixes which preserved

---

<sup>18</sup> The allomorph *we-* instead of *w-* in the latter example is conditioned by the following consonant cluster.

an old /n/ in this position, cf. Old Church Slavonic *vŭn-* ‘in-’ and *sŭn-* ‘with-’. When /n/ after /ŭ/ was dropped in closed syllables because of a regular sound change, the prefixes exhibited two forms: with /n/ before a vowel and no /n/ elsewhere (Bielfeldt, 1961: 71-72). Other prefixes adopted this behavior, resulting in forms like *vyn-* ‘out-’, from the prefix *vy-*, which originally had no /n/. The old forms here are the direct descendants of these, whereas Modern Polish has done away with this rule completely, combining all prefixes with no intermediate /n/.

The query in Table 4.5 thus retrieves many interesting phenomena in the development of verbal prefixation, but interpreting their significance requires corpus-external knowledge.

### 4.3.3 Stem Replacement with Prefix Retention

This class of verbs is the opposite of the previous class – pairs which have the same prefix, but different, unrelated stems. It is particularly interesting for the study of complex verb semantics: if the sense of a Polish complex verb is a composition of stem and prefix semantics (which is certainly not always the case), then one would expect cases where a parallel pair has different stems from different lexical roots, but the same prefix. This would be perhaps most likely in cases where a prefix is used productively and transparently, such as with motion verbs<sup>19</sup> – if a stem is replaced, the same prefixes can still be used in parallel to indicate e.g. a movement towards or away from an object. Such cases can be found by looking for items that have an identical start of string, but different strings overall. In order to find these, the opposite criteria to the previous class must be assumed: 1-4 identical characters on the left side for the prefix, and no more than 2 identical characters on the right, since, as discussed above, a minimal stem can consist of 3 characters. Admittedly, we may miss verb pairs with a coincidentally identical, longer suffix (like *-ować*, a common derivative suffix for imperfective verbs), and identical prefixes, but a different root; however, these turn out to be rare (see below).

In order to fulfill these criteria, the *LeftIdentIndex* of the pair must be between 0 and 5 and *RightIdentIndex* < 3. Unfortunately, since prefixes can be short, we must consider even verb pairs that share only a first identical character (for prefixes that are 1

---

<sup>19</sup> On Polish motion verb prefixes see Śmiech (1986).



character long), thereby risking many false positives, such as *dziać (się)*<sup>20</sup> : *dokonać (się)* ‘to happen, come to pass’, which happen to share an initial <d>, which is not a prefix. Attempting this query matches 52 pairs, of which only 28 are correct, an accuracy of only  $\approx 54\%$ . Using a list of known prefixes to match with the left side of each string (essentially an application of a greater amount of external linguistic knowledge) can cut out 10 false pairs, with no loss of correct identifications (28/42  $\approx 67\%$  accuracy). It is possible to further demand that prefixes be maximal – that is, that if a verb begins with a string that can be read as either a long prefix or a string containing a shorter prefix, the long prefix should be preferred and required in the parallel. For example in the pair *z-dumieć (się)* ‘be amazed’ : *za-niepokoić (się)* ‘become distressed’, *z-* would constitute a prefix match, but since the latter form has a longer possible maximal prefix, *za-*, this pair is (correctly) excluded as false. For maximum efficiency, the query should also regard different allomorphs of the same prefix (such as *w-/we-*, where the longer form occurs before consonant clusters, etc.) as identical. Using all these criteria, we can eliminate a further 8 false pairs with no loss of correct matches, for a final result of 28/34, or  $\approx 82\%$  precision. A manual examination of all MI3-based verb to verb correspondences has revealed only two missed cases, due to coincidentally similar suffixes and stem endings, which make the verbs appear to be related: *ze-lżyć* : *z-nieważyć* ‘insult’ and *roz-mawiać* : *roz-prawiać* ‘converse’; in both cases the verbs are not related. Recall is thus 28/30  $\approx 93\%$ , for an F-score of  $\approx 88\%$ . The retrieved records are shown in Table 4.6 on the next page.

The two groups marked in gray represent correct verb pairs that truly do share the same prefix, but have a different relationship than those in the first, main group of hits. The first three gray entries are aspectual pairs (cf. the similar group in section 4.3.2), and thus relate to a change in the grammatical category of aspect selected in either text. These hits are inevitable in at least two cases, *w-kładać* : *w-łożyć* ‘put’ and *o-glądać* : *obe-jrzeć* ‘look’<sup>21</sup>, since these pairs are suppletive, i.e. they use morphologically unrelated stems for the aspect distinction, creating the appearance of exactly the desired group of verbs. The last two entries in the table are different stem formations from the same root, and

<sup>20</sup> The reflexive particle *się* is omitted in most of the examples below where it is not essential to elucidating the parallel.

<sup>21</sup> In the latter pair *obe* (=ob before a cluster) is considered an allomorph of *o*, though their distribution is not entirely phonologically conditioned (see Swan, 2002: 282).

G lemma	W lemma	MI3	Sense
<i>pojąć</i>	<i>poślubić</i>	15.22	take/wed ( <i>see below</i> )
<i>wypchnąć</i>	<i>wyprowadzić</i>	15.08	send out
<i>zawlec</i>	<i>zdrząć</i>	13.46	take off
<i>wynosić</i>	<i>wydobywać</i>	13.37	take out
<i>rozdzielić</i>	<i>rozdwoić</i>	12.68	divide
<i>wybawić</i>	<i>wyratować</i>	12.59	rescue
<i>zapięczętować</i>	<i>zaciągać</i>	12.28	seal, shut
<i>wykować</i>	<i>wykuć</i>	12.25	dig out
<i>przypodobać</i>	<i>przyporównać</i>	11.55	compare
<i>użalić</i>	<i>ulitować</i>	11.54	pity
<i>uciszyć</i>	<i>ustać</i>	11.37	quiet, calm
<i>namazać</i>	<i>namaścić</i>	11.08	annoint
<i>opowiedzieć</i>	<i>oznajmić</i>	10.49	tell
<i>ochędożyć</i>	<i>oporzędzić</i>	10.44	clean, put in order
<i>wygnać</i>	<i>wypędzić</i>	10.41	drive out
<i>zstępować</i>	<i>schodzić</i>	10.37	descend, go down
<i>ocucić</i>	<i>obudzić</i>	10.25	rouse
<i>rozsządzić</i>	<i>rozpoznawać</i>	10.15	judge, distinguish
<i>zawrzeć</i>	<i>zamknąć</i>	10.05	close
<i>zastanowić</i>	<i>zatrzymać</i>	9.539	pause
<i>obwoływać</i>	<i>opowiadać</i>	9.031	proclaim
<i>zapuszczać</i>	<i>zarzucać</i>	8.942	cast, throw
<i>przywodzić</i>	<i>przynosić</i>	7.752	bring
<i>wkladać</i>	<i>włożyć</i>	12.61	put
<i>dotykać</i>	<i>dotknąć</i>	12.52	touch
<i>oglądać</i>	<i>obejrzeć</i>	10.35	look at
<i>porzucić</i>	<i>powiesić</i>	12.47	-
<i>obrać</i>	<i>obwieścić</i>	11.52	-
<i>pokusić</i>	<i>popaść</i>	11.25	-
<i>wiać</i>	<i>wezbrać</i>	11.2	-
<i>ubić</i>	<i>ukamienować</i>	11.08	-
<i>wypowiedzieć</i>	<i>wypełnić</i>	8.127	-
<i>zapamiętać</i>	<i>zapomnieć</i>	10.88	forget
<i>pobić</i>	<i>pozabijać</i>	9.95	kill

Tab. 4.6: Query results for verb pairs with the same prefix but a different stem.

thus belong in the next section as well (see below). The six records between the gray groups are matching errors, also exhibiting some ‘nonsensical’ prefixes, e.g. in *wiać* ‘blow’, the *w* is part of the verb stem, not a prefix. Other cases truly show the same prefix, but are not a correct parallel pair: *wy-powiedzieć* ‘declare’ and *wy-pelnić* ‘fulfill’ are not parallel “translations” of each other.

The main, first group of hits seems to confirm the working hypothesis above – the prefixes of these verbs often have a separate meaning which contributes to the meaning of the verb as a whole. Incidentally, they can often be translated with English

phrasal verbs: ‘send out’, ‘take off’, etc. Some also belong to the expected category of motion verbs: *wy-gnać* : *wy-pędzić* ‘drive out’ with the prefix signifying outward motion, but also other directed activities like *na-mazać* : *na-maćścić* ‘anoint’ with *na-* ‘on, at’, giving the sense of smearing oil ‘on someone’, or *przy-wodzić* : *przy-nosić* ‘bring, carry over’ with *przy-* signifying the sense of motion towards the destination or recipient of bringing. In more frequent prefixes, which are often used as default perfectivity markers, we also find cases of substituted verbs showing a likely coincidental choice of the same default perfectivizer: *po-jąć* ‘take’ : *po-słubić* ‘wed’ (where the former corresponds to the latter only in this context, of ‘take a woman’ = ‘marry a woman’). Here it is difficult to assign as clear a meaning to *po-* as to other prefixes, although arguably the choice of the same default prefix may be somehow related to an underlying similarity in the function or meaning of these verbs. Such ‘default’ prefix cases are difficult to identify on corpus criteria, though clearly the less frequent prefixes can be expected not to be default perfectivizers, and more likely compositional. For example *roz-* ‘apart’, a relatively uncommon prefix, has two pairs with rather transparent and meaningful prefix retention, *roz-dzielić* : *roz-dwoić* ‘divide’ (the former lemma from the root meaning ‘part’, the latter from the root of ‘two’) and *roz-sądzić* : *roz-poznawać* ‘distinguish, tell apart’ (the first lemma with the root meaning ‘judge’ and the second meaning ‘know’, with an additional prefix *po-* to create the perfective sense ‘to recognize’). In both cases the sense of division contributed by *roz-* is clear. Thus, in such semantically composite complex verbs, the stem component can be replaced by another stem with similar meaning, but the meaning contributed by the prefix (directional ‘out’ etc.) remains.

#### 4.3.4 Stem Alteration

This class contains verb pairs that are non-identical, but etymologically and morphologically related, i.e. verbs that can be derived from each other. Outside of prefixation, the most common verbal derivational means in Polish, as in other Indo-European languages, are suffixation and vowel alternation. For example, the perfective verb *przeży-ć* ‘survive’ derives its imperfective counterpart *przeży-wać* using a suffix, whereas the perfective *wróc-ić* ‘return’ has the imperfective counterpart *wrac-ać*, with a different stem vowel, as well as a different suffix. Intuitively, these pairs are likely to

have a low Levenshtein distance, but this is not accurate enough; what really makes these verbs similar are the consonant phonemes which belong to the lexical root that both verbs have in common. Focusing on consonant similarities is an approach often employed in various phonetic similarity measures, the best known of which is probably the SoundEx algorithm (for a brief description see e.g. Oakes, 1998: 130-131). One of the main principals behind SoundEx is to discard vowels (except in initial position). However SoundEx also unifies similar sounding consonants, replacing them with a numeric code (e.g. <b>, <p>, <v> and <f> are replaced by the number 1, since these graphemes represent labial sounds), whereas in this case, differences between homorganic consonants are pertinent and should be preserved.

I will therefore define a function called *PolEx* which, like SoundEx, omits vowels, but, instead of outputting an alphanumeric code for each string, outputs a human-readable string with some characters missing. The function strips the characters <a>, <e>, <i>, <o>, <u> and <y>, as well as the Polish nasal vowels <ą> and <ę>, and the additional vowel <ó>. It likewise removes the consonant <w>, which appears in many suffixes (e.g. in the example above), and also discards the infinitive endings <ć> or <c>, which occur at the end of all verbal lemmas (the latter appears in stems ending with a velar consonant). As an example, the verb *ukamienować* ‘to stone’ becomes *kmn*. Now Levenshtein distance can be measured not between the lemmas themselves, but between the simplified strings derived using this function.

Running a query for all non-identical verb pairs with a Levenshtein distance smaller than 2 between *PolEx* strings produces 72 hits, of which 20 are correct. 50 of the hits are cases of prefix changes already detected in section 4.3.2, since identical stems with a different vocalic prefix will produce identical *PolEx* strings, e.g.: *uwinąć* : *owinąć* ‘wrap’ > *n*, and similarly prefixes with the same consonant: *zadziwić* : *zdziwić* ‘marvel, amaze’ > *zdz*. Filtering the results by cross-referencing this query with the one in section 4.3.2 (or using a negative search on its criteria), we get 20/22 correct hits or  $\approx 91\%$  precision. It should be noted that this filtration removes the pair *smęcić* : *smucić* ‘mourn’, which was discussed in section 4.3.2; though the two verbs show a sort of vowel alternation, and are historically related, they are not synchronically related in Polish morphology, since the latter is a Czech loan. Thus leaving the pair out is arguably correct.

Manual examination of all verb pairs reveals that two additional pairs are missed by these criteria, which have a distance of 2 between *PolEx* strings: *pobić* : *pozabijać* ‘kill’ (the latter with an additional internal prefix *-za-*; note that this is one of the last two hits from the previous query in Table 4.6) resulting in *pb* : *pzbj*, and *objawiać* : *ujawniać* ‘reveal’ > *bj* : *jn*, where a prefix and suffix change also result in *PolEx* = 2. Recall is thus 20/22  $\approx$  91% for an F-score of  $\approx$ 91%. The results can be seen in Table 4.7.

The last two rows are errors. In the first pair *zelżyć* : *znieważyć* ‘insult’, which was missed in the last section, only the prefixes *z-/ze-* are related. The second is due to similar, long prefixes: the verbs are parallel in the text (though their meanings are somewhat different), but not related. The first four gray records are again aspectual pairs, where the different suffixes, and in some cases vowel alternations, are strategies for deriving aspectual partners. The remaining hits show various kinds of derivational morphology (note that these include the correct pair *zapamiętać* : *zapomnieć* ‘forget’, which was detected in section 4.3.3 as well). The pair *siadać* ‘sit’ : *siadywać* ‘sit (repeatedly, habitually etc.)’ shows a frequentative verb replacing a simple imperfective

G Lemma	W Lemma	MI3	PolEx G	PolEx W	LD	Sense
<i>dotykać</i>	<i>dotknąć</i>	12.519	dtk	dtkn	1	touch
<i>dopełniać</i>	<i>dopełnić</i>	12.246	dpln	dpln	0	complete
<i>rozwiązywać</i>	<i>rozwiązać</i>	11.443	rzz	rzz	0	dissolve
<i>wrócić</i>	<i>wracać</i>	9.202	rc	rc	0	return
<i>uczyć</i>	<i>nauczać</i>	14.321	cz	ncz	1	teach
<i>czuć</i>	<i>czuwać</i>	13.812	cz	cz	0	feal
<i>pogrześć</i>	<i>pogrzebać</i>	12.365	pgrzś	pgrzb	1	bury
<i>posługować</i>	<i>posługiwać</i>	12.299	psłg	psłg	0	serve
<i>zdrzemać</i>	<i>zdrzemnąć</i>	12.246	zdrzm	zdrzmn	1	slumber
<i>skarżyć</i>	<i>oskarżać</i>	12.106	skrż	skrż	0	accuse
<i>tańcować</i>	<i>tańczyć</i>	11.945	tńc	tńcz	1	dance
<i>zgodzić</i>	<i>uzgodnić</i>	11.543	zgdz	zgdn	1	agree
<i>ukamionować</i>	<i>ukamienować</i>	11.079	kmn	kmn	0	stone
<i>zapamiętać</i>	<i>zapomnieć</i>	10.883	zpmt	zpmn	1	forget
<i>pełnić</i>	<i>spełniać</i>	10.575	płn	spłn	1	fulfill
<i>czekać</i>	<i>oczekiwać</i>	10.401	czk	czk	0	wait
<i>dziwować</i>	<i>dziwić</i>	10.329	dz	dz	0	marvel, amaze
<i>umywać</i>	<i>myć</i>	10.253	m	m	0	wash
<i>siadać</i>	<i>siadywać</i>	8.895	sd	sd	0	sit
<i>zabieżeć</i>	<i>zabiec</i>	8.825	zbż	zb	1	run across
<i>zelżyć</i>	<i>znieważyć</i>	13.216	zlż	znż	1	insult
<i>przewieźć</i>	<i>przybyć</i>	9.157	przż	przb	1	cross/arrive

Tab. 4.7: Query results for verb pairs with stem alterations.

as a sort of explicitation<sup>22</sup>. Two pairs show fluctuation between verb formation with a thematic (i.e. with an initial vowel) or an athematic suffix (just the /ć/ ending, with no vowel): athematic *pogrześć* : thematic *pogrzebać* ‘bury’ (in the former form, which has gone out of use, the /b/ of the stem -*grzeb-* is assimilated to /ś/ by the infinitive /ć/) and in the other direction, *zabieżeć* : *zabiec* ‘run across’ (in the former, now defunct form, an underlying /g/ at the end of the stem -*bieg-* is palatalized into /ż/, in the latter it fuses with the infinitive ending into /c/). There is also fluctuation between using simple imperfective verbs and secondary imperfectives, derived from a prefixed perfective verb: *uczyć* (simple) : *na-uczać* (complex) ‘teach’, *pełnić* (simple) : *s-pełniać* (complex) ‘fulfill’ etc., and vice versa *u-mywać* (complex) : *myć* (simple) ‘wash’. The tension between marking imperfectives with suffixes or using simplex verbs is most obvious in a case showing imperfectivizing suffixation of a stem that was already an imperfective simplex: in *dziwować* : *dziwić* ‘marvel, amaze’, the former form is no longer in use, perhaps because it is marked as imperfective ‘twice’ – by lack of a prefix, and the imperfective suffix *ować#*. The imperfective suffix’s form has sometimes changed too, e.g. *posługować* : *posługiwać*, with *R4ywać#* (realized as <i>wać</i>) replacing *ować#* (both suffixes are in use for different verbs in Modern Polish, but the former form no longer exists for this verb).

#### 4.3.5 Total Substitution

The last type of verb-verb pairs involves two completely unrelated lemmas, sharing no common parts. To find these, results for verb-verb correspondences must be cross-referenced with the results from the previous classes’ queries, in order to filter these out (alternatively, a search for verb pairs with high Levenshtein distance could produce similar results). The biggest problem is that results will also include nearly all erroneous pairs (i.e. incorrect parallels), since these are likely to be mostly very dissimilar verb pairs, which look exactly like the case of a change to an unrelated lemma. At present, no automatic method has been found to distinguish these: out of 159 query hits, only 30 have been manually determined to be correct parallel pairs, and these are shown in Table 4.8.

---

<sup>22</sup> The phenomenon of explicitation, i.e. the replacement of a more polysemous or wider-sensed term with a more specific one, is often observed in secondary versions of texts, and especially in translations (see Hansen-Schirra and Teich, to appear, and references there).

G Lemma	W Lemma	MI3	Sense	G Lemma	W Lemma	MI3	Sense
<i>wadzić</i>	<i>spierać</i>	15.54	quarrel	<i>baczyć</i>	<i>dostrzegać</i>	10.84	notice
<i>wykorzenieć</i>	<i>powyrywać</i>	15.00	uproot	<i>dopuszczać</i>	<i>pozwolić</i>	10.79	allow
<i>zawisnąć</i>	<i>opierać</i>	12.47	depend, lean	<i>chwiać</i>	<i>kiwać</i>	10.69	shake, nod
<i>zaniechać</i>	<i>ustąpić</i>	12.25	give up	<i>gwałcić</i>	<i>naruszać</i>	10.69	violate
<i>dziać</i>	<i>dokonać</i>	12.10	happen	<i>naprawić</i>	<i>odnowić</i>	10.61	repair, renew
<i>pomdleć</i>	<i>zasłabnąć</i>	11.98	faint	<i>dowiadawać</i>	<i>wypytywać</i>	10.37	find out, investigate
<i>prawować</i>	<i>procesować</i>	11.69	sue	<i>obwarować</i>	<i>zabezpieczyć</i>	10.36	guard
<i>zamilknąć</i>	<i>oniemieć</i>	11.52	be silent	<i>przechadzać</i>	<i>wędrować</i>	10.15	wander
<i>odziedziczyć</i>	<i>posiąść</i>	11.37	inherit, possess	<i>odmładzać</i>	<i>mięknąć</i>	10.05	become young/soft
<i>cierpieć</i>	<i>znosić</i>	11.25	suffer	<i>spodzierać</i>	<i>domyślać</i>	9.752	expect, suppose
<i>przymawiać</i>	<i>grozić</i>	11.22	call/threaten	<i>robić</i>	<i>pracować</i>	9.725	work
<i>rozgniewać</i>	<i>oburzyć</i>	11.21	anger	<i>złęknąć</i>	<i>przerazić</i>	9.706	frighten, fear
<i>pozabijać</i>	<i>mordować</i>	11.11	murder	<i>ukusić</i>	<i>zaznać</i>	8.757	experience
<i>patrzeć</i>	<i>baczyć</i>	11.10	look	<i>poprzysięgać</i>	<i>zaklinać</i>	8.381	curse
<i>wstydząć</i>	<i>uszanować</i>	10.99	be ashamed	<i>lśnić</i>	<i>zajaśniać</i>	8.379	shine

**Tab. 4.8: Unrelated verb-verb pairs.**

We can also add to these results the pair *zmiłować* : *zlitować* ‘pity’ which was falsely detected in section 4.3.2 as a case of prefix change, and *przewieźć* ‘cross’ : *przybyć* ‘arrive’, which was detected in the previous section as a stem alteration. In most cases found in the table, the older lemma is still usable in Modern Polish, though not always: e.g. *ukusić* ‘experience’ and *spodzierać* ‘expect’ are not listed in the PWN online dictionary (<http://www.pwn.pl/>, the largest online Polish dictionary) while *wadzić* ‘quarrel’ is noted as an archaism; the more basic and colloquial dictionary used by the tagger (see section 2.3) further had to be expanded with *przymawiać* ‘call’, *odmładzać* ‘become young’, *pomdleć* ‘faint’, *poprzysięgać* ‘curse’ and *obwarować* ‘guard’, which were all unlisted. In cases where the older lemma exists in Modern Polish, the sense required by the text is sometimes given by a less ambiguous word (i.e. explication): *robić* : *pracować* ‘work’, where the older lemma currently means more generally ‘do, make’, or *chwiać* : *kiwać* ‘nod (one’s head)’, where *chwiać* means generally ‘to shake’. In some cases the text has been phrased differently, leading to real differences in meaning, e.g.: *przymawiać* ‘call’ : *grozić* ‘threaten’, with the contents of the call (a threat) already included in the verb.

#### 4.3.6 Non-Verbal Lemmas

In this class we can expect to find verbal lemmas which correspond to items with lemmas that either have different parts of speech, such as nouns or adjectives, or even collocations, possibly containing verbal lemmas, which are however different from



G Lemma	Sense	W Lemma	Sense	MI3
<i>spłodzić</i>	to beget	<i>ojciec</i>	father	18.569
<i>pokłonić</i>	to honor	<i>pokłon</i>	honor	14.019
<i>nauczyć</i>	to teach	<i>uczony</i>	scholar	13.905
<i>urzezać</i>	to castrate	<i>trzebień</i>	eunuch	13.579
<i>naradzić</i>	to deliberate	<i>narada</i>	deliberation	12.632
<i>próżnować</i>	to be idle	<i>bezczyunny</i>	idle	12.558
<i>policzkować</i>	to slap	<i>pięść</i>	fist	11.883
<i>załatać</i>	to patch	<i>łata</i>	patch	11.861
<i>odpocząć</i>	to relax	<i>ukojenie</i>	relief	11.799
<i>rachować</i>	to reckon	<i>obrachunek</i>	reckoning	11.293
<i>nazowieć</i>	to name	<i>nadać imię</i>	to give a name	11.009
<i>dokonać</i>	to accomplish	<i>koniec</i>	end	10.936
<i>pośmiewać</i>	to ridicule	<i>pośmiewisko</i>	laughing stock	10.539
<i>przywrócić</i>	to restore	<i>znow</i>	again, anew	10.46
<i>zdumieć</i>	to amaze	<i>zdumiony</i>	amazed	10.106
<i>klaniać</i>	to bow down	<i>pokłon</i>	honor	8.431

**Tab. 4.9: Verbal lemmas corresponding to non-verbal lemmas.**

simple verbs. Finding these can be accomplished easily by searching for lemmas not matching any verbal part of speech (which subsumes collocations as well). However, it is again impossible to distinguish between true matches and matching errors (in fact, errors may be even more likely in verb-non verb pairs, since a verb-verb match is *a priori* likelier than a match with another part-of-speech). Table 4.9 show the 16 correct matches manually sorted out of a total of 83 hits.

Remarkably, all but one correspondence are also single lemmas: only the verb *nazowieć* ‘name’ is paralleled by the collocation *nadać imię* ‘give a name’. This is most likely because slight variability in collocations makes their association with a parallel lemma split up between various possibilities, while a core lemma involved in several phrasings can achieve a higher MI3 score. For example, the verb *pokłonić* ‘to honor’ appears to have the parallel noun from the same root, viz. *pokłon* ‘honor’. However in the text this is always part of a larger phrase, usually either *oddać pokłon* ‘give honor’ or *złożyć pokłon* ‘pay honor’ (lit. ‘lay down honor’). Since *pokłonić* predicts the appearance of a parallel *pokłon* better than either one of the two collocations, the noun is found to be the best match. Such verb-noun pairs, especially from the same root, are the most common case in this class: *naradzić* ‘to deliberate’ : *narada* ‘deliberation’, *rachować* ‘to reckon’ : *obrachunek* ‘reckoning’, *pośmiewać* ‘to ridicule’ : *pośmiewisko* ‘laughing stock’ etc. One of the most interesting cases is the pair *spłodzić* ‘beget’ : *ojciec* ‘father’,



which is very significant on the strength of evidence from the repetitive structures in the genealogy of Jesus in Matthew 1 ('X begat Y'). Much like in English, the archaic 'beget' verb was abandoned, replaced instead by the construction 'X was the father of Y' (in Polish *X był ojcem Y*). Since I have examined lemmas, and not grammatical categories, the various proper names in the positions of X and Y do not collocate, and the very common *był* 'was' does not either, leaving a very significant correspondence between the verb and the word for 'father'. In a sense, however, this is a completely legitimate and correct parallel (the lexical meaning of 'beget' is really rendered by 'father' here), and a finer mapping, e.g. including the copula verb, is not possible without considering grammatical categories (see next chapter).

Less frequent are correspondences with adjectives, and these are more questionable too, since the adjectives usually parallel a participial form. In the case of *próżnować* 'to be idle' : *bezczynny* 'idle' the parallelism stems from the fact that this verb consistently appears in a participial form in GMat: *próżnujący* 'idling'. The WMat form is in this case an adjective replacing a participle. However, in the case of *zdumieć* 'to amaze' : *zdumiony* 'amazed', the latter form is simply the participle of the former's verbal lemma; the only reason this pair was not recognized as an identical verb-verb lemma correspondence in section 4.3.1, is that the lemma *zdumiony* appears in a separate entry in the tagger's dictionary with the part-of-speech adjective, and the tagger chose this option over a productively derived passive participle. In other words, the lexicon used to tag the corpus is skewing results to fit its contents. The same can be said for the pair *nauczyć* 'to teach' : *uczony* 'scholar', where the Modern Polish word *uczony* is in fact a lexicalized passive participle ('taught' > 'scholar', cf. the English adjective *learned* with the sense 'educated, intelligent'). The older lemma appears in the text with the form *nauczony* (the exact same formation, but with a prefix), but since it is not in the lexicon, it is considered to be productively derived from the verb *nauczyć* 'teach'.

In only one case is there a parallel with an adverb, *przywrócić* 'to restore' : *znów* 'again, anew', which is again part of a longer phrase: *stać się znów* 'to become again'. Since *stać się* 'become' is quite common, the adverb produces a more significant match than the verb.

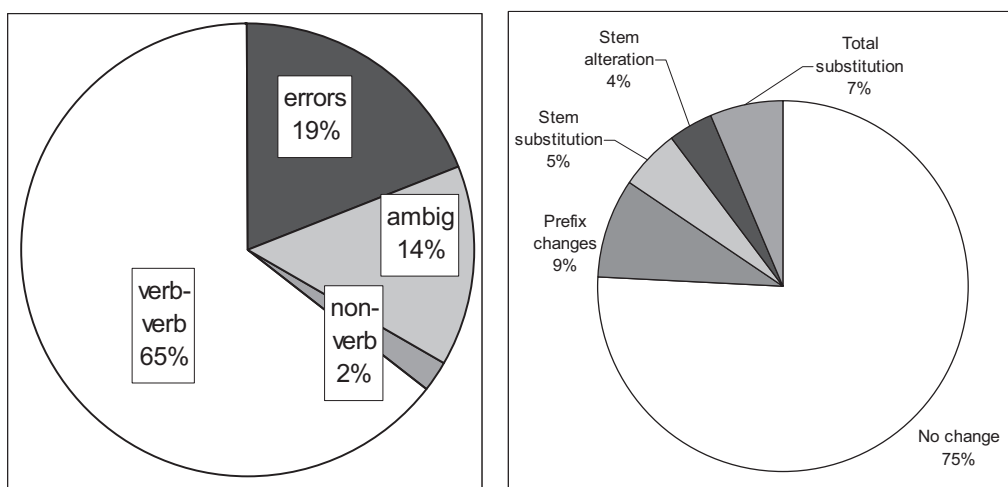
## 4.4 Summary and Evaluation

In this chapter I have examined correspondences between verbal lemmas across the two corpora, extracting and classifying the results mostly automatically. The application of SQL queries combined with string similarity measures has proven to be a good tool for extrapolating pairs of identical corresponding verbs, or ones differing only in prefixes, the stem's root, or stem formation, directly from the part-of-speech-tagged and lemmatized corpus data. The results made possible by these techniques give new dimensions for historical insight, in a quantitative way generally not provided by traditional historical grammars.

Out of a total of 760 verbal lemmas in GMat it has been possible to find correct unambiguous matches for a total of 507, which have been detailed in the previous sub-sections. For the remaining lemmas, the best parallel was erroneous (146 cases), or else multiple parallels received the same scores (107 cases), either because of polysemy (or at least stylistic variation producing multiple parallels), or because so few attestations were available that the association measure could not resolve to which of several rare elements in the parallel section an item corresponds. This situation is partly dependent on corpus size, though it is expected that even much larger parallel corpora will contain a substantial portion of rare lemmas for which 'consistent parallels' (from an association measure point of view) cannot be established, due to Zipf's law: there will always be many unique *hapax legomena*, fewer *dis legomena* etc. Evaluating precision and recall for the matching process itself is difficult: a recall of 760/760 pairs cannot be realistically expected, since multiple matches cannot necessarily be resolved to a single match by a human either – some items truly have different, correct matches. However considering error cases, precision is no higher than  $507/637 \approx 80\%$ , i.e. the ratio of correct matches to unambiguous matches claimed by MI3.

The proportion of verb-verb pairs among the correct results is very high, as can be seen on the left in Figure 4.3. This can mainly be ascribed to the extremely faithful parallelism between the texts, which adhere to the languages from which they were translated (and, presumably, especially to the Latin of the vulgate), and also to their internal parallelism, since the translators of the Warsaw Bible also had the Gdansk Bible available to them. More surprising is perhaps the rarity of collocation parallels observed

in section 4.3.6. However as already mentioned, this is partly an artifact of the association measure, which in effect admits only the most robust collocations, i.e. in cases where their components' frequencies make them insignificant in isolation (e.g. the word for 'name' is common outside the collocation 'to give a name', and this component lemma alone is therefore not significantly associated with the verb 'to name'). Weighting the significance of collocations (since multiple items are likely to appear less often than their components) might produce different results.



**Fig. 4.3: Verb correspondences and distribution of verb replacement types.**

Internally, the verb-verb match group is dominated by unchanged lemma pairs (including non-identical aspectual partners), taking up about three quarters of all unambiguous pairs (the graph on the right of Figure 4.3). Even if we recall that in case of multiple matches, a pair of identical lemmas is given priority (cf. section 4.3), the proportion of verbs that have remained unchanged is very high. Moreover, where a verb has been replaced, more often than not a part of the old verb remains: the largest group has only a prefix change, and total substitution is not very common. This means that Poles today can read the Gdansk Bible with relatively little lexical difficulty, which fits well with the fact that the Gdansk Bible remained in use for Protestant Poles well into the 20<sup>th</sup> century.

The relatively strong presence of partial replacements offers empirical evidence in support of the compositionality of complex verbs across time – prefixes, stems and roots

can change independently, especially in cases of productive, transparent prefixation (among other cases in motion verbs, cf. section 4.3.3; on their central role in the development of Polish prefixation see Śmiech, 1986). These results also raise questions for future studies which could compare the proportions of partial change types to those in other parallel texts in different time spans and languages. One might imagine, for instance, that prefixes may be more or less lexicalized and therefore more or less open to independent change in different Slavic languages, or Indo-European ones (cf. the more lexicalized prefixes of Romance verbs).

There are however many factors that are still disregarded in this representation. Firstly, as already mentioned, both polysemy and stylistic variation mean that verbs generally have more than one parallel, albeit perhaps less significant than a dominant one. In particular, I have not examined different developments for active versus reflexive variants of the same verb. This would be difficult to achieve in an unparsed corpus, since Polish reflexive verbs are marked by the reflexive clitic *się*, which may appear either before or after the verb, with other clitics possibly intervening. Its automatic association to one verb in a sentence is also non-trivial. In general, developing a more complex model of multiple alternative context-dependent substitutions, factoring in the reliability of their association, is an interesting and desirable future objective, which may however require a more intricately-annotated corpus, with e.g. syntactic and sense annotation. In fact, establishing what proportion of lemmas are differentiated into different lemmas for different senses is in itself interesting, since it has been suggested that the appearance of more such distinctions is related to the semantic distance between languages (cf. e.g. Resnik and Yarowsky, 2000, who draw attention to the fact that parallel texts in more distant languages pairs produce better sense disambiguation by comparing variant translations). This could then be used as a quantifiable measure to judge such distances, though its usefulness would require an independent evaluation.

It is furthermore problematic to accept MI3 as a gold standard for finding parallels. Precision and recall rates given in the previous sections are for the proportion of correctly retrieved verbs within the MI3-based concordance. Although MI3 has proven fairly successful for this purpose, it by no means recovers all parallels as a human would manually extract them, but only those which show an unambiguous association with a

parallel item. Thus, while this may be viewed as positive (only well-attested, clear cut pairs are retrieved, without subjective interpretation), it must be kept in mind that the recalled items subsume an association defined in MI3 terms, which may be inaccurate with respect to a human alignment-based gold standard. On the other hand, human alignment, not to mention human judgment on what a consistent alignment is, can also be expected to introduce some inaccuracy (or simply variance). In the end, the question of what should constitute a diachronic correspondence could be debated on theoretical terms at great length, but it is perhaps not entirely solvable for all cases, and almost certainly not on purely statistical grounds.

Finally, although the retrieved pairs can be classified using string comparison functions in most cases with good success (F-scores around 90%), it remains clear that results can only be interpreted with knowledge external to the corpus. There is for example no way for a query to formally distinguish between normal prefix changes and the exceptional case of the forms *wn* and *w* ‘in’, which is due to a morphotactic rule change (cf. section 4.3.2). This is also true regarding the nature of ‘replacements’, since as already mentioned, a recognized pair does not mean that one form completely replaced the other over time, only that one was chosen over the other in a modern text in a parallel context. For this last difficulty, the opportunity presents itself to also use larger, non-parallel historical corpora, against which one could compare results from the parallel corpus, and find out whether or not forms are in synchronic variation, or else have gone out of use.

## 5 Syntactic and Grammatical Change

The same principles applied to the study of lexical items in the previous chapter can also be adapted to the study of changes in grammatical categories and syntactic constructions. Syntactic or grammatical change is understood here more broadly than just changes in word order, which have been less significant for the development of Polish with its relatively free word order than for the more rigid Germanic or Romance languages. Rather, I will be interested in any consistent correlations between syntagms and grammatical annotation layers here.

It is possible to look for significant correlations between parallel constructions by considering different kinds of items other than lemmas, such as part-of-speech tags. However, while single token grammatical categories like finite verbs can be identified easily in this way, longer constructions will have to be defined in terms of flat, recurring patterns of tokens, since the corpus is not parsed. This level of abstraction is not ideal, but the rich case system in Polish often makes establishing subject, object, congruent attributes etc. possible even without a parse. In principle, however, a parsed corpus could be used to identify structures more accurately, and their occurrences in aligned sections could be correlated in the same way (more on this in chapter 6).

The next section will briefly describe the procedure used to remove lexical information from tokens to allow for better generalizations. The following three sections explore different correspondences between the corpora that can be found using these delexicalized tokens: section 5.2 investigates the replacement of indeclinable predicative adverbial participles with finite verbs; section 5.3 looks at the decline of possessive adjectives in favor of the nominal genitive; and section 5.4 describes changes in the copula verbs used to form passive predications with passive participles and their interaction with verbal aspect. The potential and limitations of some of the methods used in these sections will be discussed within the summary and evaluation in section 5.5.

### 5.1 Reduction of Token Sequences

In order to pair grammatical items between the corpora, it is necessary to abstract away lexical information, such as the strings marking lemmas and word forms, effectively

decreasing the amount of token types for better generalization power. Therefore instead of looking at each occurrence of a lemma like *kapłan* ‘priest’, or one of its inflected forms, I will regard it as an occurrence of the category ‘noun’, grouped together with other nouns. On the other hand, it might be useful to retain the lemmas for certain classes of words which play important grammatical roles, such as prepositions, or very common ‘function word’ lemmas such as *być* ‘to be’, since the exact identities of these can distinguish different grammatical structures, and are much less interchangeable within a part-of-speech category. For the purposes of the studies in this chapter I have decided to discard lexical information (i.e. lemmas and word forms) for all tokens having a verb, noun, adjective or adverb as a lemma, with the exception of lemma types having a frequency of over 1% in the corpus. This leaves a heuristic class of ‘function words’ which includes prepositions, conjunctions and pronouns, as well as frequent lemmas like the aforementioned *być* ‘to be’.

It is further possible to omit grammatical categories with a lexical nature within the annotation of each token. For example gender is probably less pertinent for many syntactic questions than, say, grammatical case. The same applies to number information: most grammatical structures (e.g. a transitive verb frame) accommodate singular and plural nouns, verbs etc. Figure 5.1 gives an example of an abstracted token trigram. The lemmas *być* ‘to be’ and *na* ‘on’ are not stripped since they are very frequent, and in the case of *na* belong to a reserved part-of-speech class (prepositions). The more ‘lexical’ lemma *pustynia* ‘desert’ is stripped away, as are its gender and number.

IMPFV PRES 3 SG	PREP	LOC SG F		
jest	na	pustyni	>	[VFin być impfv pres 3] [Prep na] [S loc]
<i>he is in the desert</i> (WMat 24:26)				

**Fig. 5.1: Abstracting a token sequence.**

Finally, I will be using a different tokenization in this chapter. Unlike the previous chapter, which was concerned with lemmas, this chapter will focus on grammatical categories such as finite verbs. However as discussed in section 2.2, tokenization of these grammatical categories is not always trivial on account of the behavior of clitics. For this reasons, the second tokenization with ‘link tokens’ will be used in this chapter, allowing for example both separable multi-token constructions and inseparable (fused or clitic-

less) single-token constructions to equally stand for e.g. a past tense or conditional finite verb. The link tokens will make the preparation of a concordance of all occurrences of such categories possible in the same way as before.

## 5.2 Decline of the Active Past Participle

As a simple example of using grammatical categories for pair matching, I will examine correlates of the active past participle (PPA for short), also known as the ‘adverbial participle of prior action’ or simply ‘past gerund’ (Swan, 2002: 301-302). This indeclinable participle, which is formed only from perfective verb stems with the suffix *szy#* (or *wszy#* after a vowel), expresses a predicate which takes place before the action of another, inflected verb, with the same subject, e.g.:

- (10) I      **wszedłszy** w dom,                      znaleźli                      dzieciątko z Maryją  
          and enter-PPA in house found-VFIN-PFV-PAST-3-PL-V child with Mary  
          *And having entered the house, they found the child with Mary* (GMat 2:11)

The form is all but obsolete in Modern Polish, appearing ever more rarely in only the most literary contexts (ibid.).

In order to find out what happens to this form in WMat, I will regard occurrences of the corresponding part-of-speech tag (VPartPastAct) as target items. After preparing a verse-aligned concordance of all tokens stripped down according to the guidelines outlined in the previous section, I group and count pairs featuring the VPartPastAct tag, and calculate MI3 for each possible pair, in exactly the same way as lemma correspondences in chapter 4. The best result for this category can be seen in row 1 of Table 5.1, and contrasted with the null hypothesis, that the category is retained, in row 2.

Row	GTok	WTok	A	B	C	MI3
1	[VPartPastAct]	[VFin pfv past 3]	3502	12643	340	19.786
2	[VPartPastAct]	[VPartPastAct]	3502	1215	72	16.447

**Tab. 5.1: Results of best match and no change pairs for PPA in GMat.**

As row 2 shows the active past participle is still used in WMat in parallel, but the most significant WMat correlate of the form in GMat is a perfective, 3<sup>rd</sup> person past tense verb,



the most common narrative form. This parallel corresponds to the aforementioned decline of the active past participle in Modern Polish. Its replacement by inflected, perfective past tense forms can be seen in example (11), whereas the less frequent case of its retention is illustrated in example (12) (gender, number etc. are omitted for the non-highlighted tokens to save space):

(11) PPA : VFin

G: A **posławszy** je do Betlehemu, rzekł  
and send-PPA them to Bethlehem said-VFIN  
W: I **posłał** ich do Betlejem i rzekł  
and sent-VFIN-PFV-PAST-3-SG-M them to Bethlehem and said-VFIN  
*And he sent them to Bethlehem and said* (Mat. 2:8)

(12) PPA : PPA (no change)

G: A **opuściwszy** Nazaret, przyszedł, i mieszkał w Kapernaum  
and leave-PPA Nazareth came-VFIN and dwelt-VFIN in Capernaum  
W: I **opuściwszy** Nazaret, przyszedł i zamieszkał w Kafarnaum  
and leave-PPA Nazareth came-VFIN and dwelt-VFIN in Capernaum  
*And after leaving Capernaum, he came and dwelt in Capernaum* (Mat. 4:13)

In this way the meaningful correlation between old participles and new finite verbs can be detected, despite the overall high frequency of finite verbs and the low frequency of PPAs in WMat. Too few cases are retained to make the mutual informativity of the PPA-PPA match more significant than the finite verb match: only 75 PPA's are found in WMat, compared to 241 in GMat, and while almost all WMat cases are paralleled by a GMat PPA (cf. the 72 cooccurrences in row 2, column C above), the other GMat cases are almost all rendered by a finite verb as in example (11).

### 5.3 Possessive Adjectives versus Genitival Possession

Up until this point I have examined only single tokens. It is however also possible to examine longer syntagms using the same means, if their occurrence can be operationalized in the form of a verse-aligned concordance (i.e. a list of how often each syntagm appears in each verse in each corpus). In this section I will examine token

sequences to detect changes in the usage of possessive adjectives. These adjectives are derived from proper nouns, most often with the suffix *owy#*, and were used in Old and Middle Polish, just as already in the oldest Slavic documents, to express possession (Pisarkowa, 1984: 128-129; Rospond, 2003: 195), e.g. *Syn Dawidowy* ‘Son of David’, literally: ‘Davidian son’.

In order to examine this construction, I will refine the token abstracting procedure described in section 5.1 with a function to identify grammatical agreement (i.e. congruence in case, number, gender and person). This function receives the grammatical analyses of all tokens in a sequence before gender and number information is discarded, and appends the string *agr* (for ‘agreement’) for any item that may be congruent. Once possible congruences have been established number and gender information can be discarded. Figure 5.2 shows an example of the output of the abstraction process with the *agr* function. The tokens *dobre* ‘good’ and *nasienie* ‘seed’ are stripped, but receive the feature *agr*, since they are congruent (the function itself is given in appendix B for reference).

IMPFV PRES 3 SG	ACC SG N	ACC SG N	
sieje	dobre	nasienie	> [VFin impfv pres 3] [Adj acc <b>agr</b> ] [S acc <b>agr</b> ]
sows	the good	seed	(WMat 13:37)

**Fig. 5.2: Stripped token sequence with the *agr* function.**

Using this function it is now possible to identify possessive adjectives which are in congruence with a noun. The 5 best MI3 results of a query for parallel bigrams with a congruent noun and its possessive adjective in GMat are shown at the top of Table 5.2.

Row	GMat	WMat	A	B	C	MI3
1	[S gen agr] [AdjPos gen agr]	[S gen agr] [AdjPos gen agr]	142	72	8	13.39
2	[S voc agr] [AdjPos voc agr]	[S voc] [SN gen]	97	77	5	11.81
3	[S nom agr] [AdjPos nom agr]	[S nom] [SN gen]	42	199	6	12.43
4	[S dat agr] [AdjPos dat agr]	[S dat agr] [AdjPos dat agr]	33	31	2	10.71
5	[S acc agr] [AdjPos acc agr]	[S acc] [SN gen]	63	194	3	8.88
6	[S gen agr] [AdjPos gen agr]	[S gen] [SN gen]	142	207	3	7.62
7	[S gen agr] [AdjPos gen agr]	[SN gen] [SN gen]	142	71	2	7.41
8	Total: [S* agr] [AdjPos agr]	Total: [S* nom] [SN gen]	4779	2169	31	9.26
9	Total: [S* agr] [AdjPos agr]	Total: [S* agr] [AdjPos agr]	2169	696	14	8.60

**Tab. 5.2: Parallel bigrams with a congruent possessive adjective in GMat.**

As the query results in rows 2, 3 and 5 show, the construction was often replaced by qualifying the noun (POS-tag S) with a proper noun (SN) in the genitive (e.g. genitival ‘David’s son’ replacing adjectival ‘Davidic son’), a phenomenon which gradually reduced use of the old construction beginning as early as the 16<sup>th</sup> century (Rospond, 2003: 195). Since the stripping function does not discard case information, there are separate entries for different cases of the same construction. In some grammatical cases, the adjectival construction has been left intact more often, e.g. in rows 1 (genitive case) and 4 (dative case), which have identical pairs. However, the next best match for the construction in row 1 would also be formed by the noun + genitive proper noun construction if it had not been split between two versions (see the gray rows in the table): one type of sequence qualifies a proper noun (SN) with another proper noun and the other a common noun (S), and they are counted separately. An examination of all cases and noun types together (last two rows) shows that in the two best overall matches, the association between the old possessive adjective and the newer genitive construction is in fact not much stronger than the archaic adjective-adjective correspondence, meaning many cases were not replaced (in fact almost a third, as revealed by column C in the last two rows and verified manually in the text). This can probably be ascribed to the relative conservatism of the text. Examples (13) and (14) illustrate both types of parallel:

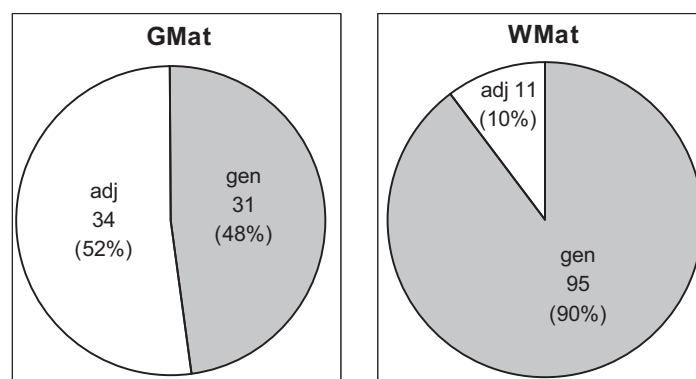
(13) Possessive adjective : possessive adjective (no change)

G: Wtedy przystąpiła do niego matka **synów** **Zebedeuszowych**  
 then approached to him mother sons-GEN Zebedee-ADJPOS-GEN  
 W: Tedy przystąpiła do niego matka **synów** **Zebedeuszowych**  
 then approached to him mother sons-GEN Zebedee-ADJPOS-GEN  
*Then the mother of the sons of Zebedee approached him* (Mat. 20:20)

(14) Possessive adjective : possessive genitive

G: I pełni się w nich **proroctwo** **Izajaszowe**  
 and fulfill REFL in them prophecy-NOM Isaiah-ADJPOS-NOM  
 W: I spełnia się na nich **proroctwo** **Izajasza**  
 and fulfill REFL on them prophecy-NOM Isaiah-SN-GEN  
*And in them the prophecy of Isaiah is fulfilled* (Mat. 13:14)

The accurate alignment of the corpus allows the correct identification of the old construction and its competition with its younger contender, despite the relative infrequency of the phenomenon. How powerful the possibilities granted by the parallel text are can be gleaned from a comparison with the purely frequency-based alternative. A simple query on the proportion of the two competing constructions between corpora (disregarding parallelism) would show the genitive construction to have become much more dominant than it actually is in this use (Figure 5.3).



**Fig. 5.3: Proportions of the proper noun-genitive and possessive adjective constructions.**

This is because one cannot guarantee that all occurrences of the genitive construction in the new corpus are in fact translating old possessive adjectives; they may simply represent a coincidental appearance of a genitive proper noun next to another noun, in a context which never housed a possessive adjective. Indeed, the construction appears in WMat almost 3 times as often as there even are possessive adjectives in GMat, and about 150% more often than the possessive adjectives and the genitive construction in GMat put together. Some matches are therefore clearly unrelated to this development, and this can only be discerned by taking advantage of the parallel alignment as in Table 5.2.

## 5.4 Passive Participle Copulas

In the last two sections I concentrated on using delexicalized sequences of tokens, completely disregarding lemmas. In this section I will attempt to combine lemma correspondences, similar to those extracted in chapter 4, with grammatical category-based querying as in the last two sections, in order to examine the development of passive

predicative syntagms comprised of copulas and passive participles. The question I will be interested in here specifically is which copula verb is used with these participles. In order to examine the parallel correlates of passive participles, I will use part-of-speech information to find their occurrences, ignoring lemmas as described in section 5.1. I will then create a parallel concordance of the lemmas occurring adjacently to the participles, similarly to the concordances in chapter 4. Searching for adjacent tokens both before and after participles is necessary because of the relatively free word order in Polish: predicative passive participles may be either preceded or followed by the copula. Although the participles are not always (though usually) adjacent to the copula, I will consciously omit non-adjacent cases, since without a parsed corpus it is otherwise very difficult to distinguish the rare, non copula-adjacent predicative passive participles from the more common attributive or nominalized passive participles, which can coincidentally appear in a sentence that contains a copula.

Table 5.3 shows the results of a query searching for correlated parallel lexical lemmas (i.e. excluding punctuation, conjunctions etc.) that occur in all bigrams containing passive participle tags. The top two matches are relatively frequent and have a substantial MI3 score. These are verb-verb pairs (interestingly despite the fact that a verbal POS tag was not specified in the query), although a third matching verb-pair can be found with a much lower score and only 4 cooccurrences (column C). These are also the copulas that appear in both corpora with passive participles.

Row	GMat	WMat	A	B	C	MI3
1	<i>być</i>	<i>być</i>	188	164	80	13.07
2	<i>być</i>	<i>zostać</i>	188	31	18	9.018
3	<i>bywać</i>	<i>być</i>	10	164	4	4.337

**Tab. 5.3: Correlated lemmas next to passive participles.**

Row 1 shows that passive participles usually occur next to the verb *być* ‘be’ (‘to be verbed’), but row 2 shows another parallel, with *zostać* ‘become’, which is used regularly as a copula with perfective passive participles (in the perfective future or past ‘will be/was verbed’). The third row shows the less common frequentative form of the verb ‘to be’, *bywać* ‘to be (often, repeatedly)’, which is used in GMat occasionally to express iterative or habitual passives, but does not appear in WMat.

Although the use of *zostać* is a much younger development in the prehistory of Polish (*być* is the older form found in all Slavic languages), both forms were in use throughout the recorded history of Polish and evolved alongside the establishment of verb-stem aspect (Długosz-Kurczabowa and Dubisz, 2006: 316). Yet this construction does not occur in GMat at all, where the lemma *zostać* is used as the perfective form for the fully lexical verb ‘to remain, stay’, and *być* is used with both aspects of passive participles. In the modern language *zostać* is prevalent for perfective participles, though use of *być* continues both with imperfective participles to express imperfective actional passives (‘is being/was being/is going to be verbed’) and occasionally with perfective participles to express a statal or resultative sense of completion (approximately ‘had/has/will have been verbed’) (see Swan, 2002: 312-314). Some parallel examples from the corpora illustrate the differences in usage:

(15) Imperfective + *być* : imperfective + *być* (no change)

G: I        **byli**                    **chrzczeni**                    od niego w Jordanie  
          and be-PAST   baptized-PASS-IMPFV   by   him   in   Jordan  
 W: I        **byli**                    **chrzczeni**                    przezeń w rzece Jordanie  
          and be-PAST   baptized-PASS-IMPFV   by-him   in   river   Jordan  
*And they would be baptized by him in the river Jordan* (Mat. 3:6)

(16) Perfective + *być* : perfective + *być* (no change)

G: i            **uzdrowiona**            **była**    niewiasta    od    onej    godziny  
          and    healed-PASS-PFV   be-PAST    woman    from    that    hour  
 W: I        od    tej    chwili    niewiasta    **była**            **uzdrowiona**  
          and    from    this    moment    woman    be-PAST    healed-PASS-PFV  
*And from that moment the woman was healed* (Mat. 9:22)

(17) Perfective + *być* : perfective + *zostać*

G: gdy    Maryja, matka jego,    **poślubiona**                    **była**    Józefowi  
          when    Mary    mother    his    married-PASS-PFV   be-PAST    to-Joseph  
 W: Gdy    matka jego, Maria,    **została**                    **poślubiona**                    Józefowi  
          when    mother    his    Mary    become-PAST    married-PASS-PFV   to-Joseph  
*When his mother Mary was betrothed to Joseph* (Mat. 1:18)

(18) Imperfective + *bywać* : imperfective + *być*

G: I ubogim Ewangelija **opowiadana** **bywa**  
 and poor gospel tell-PASS-IMPV be-FREQ-PRES  
 W: a ubogim **zwiastowana** **jest** ewangelia  
 and poor tell-PASS-IMPV be<sup>23</sup>-PRES gospel  
*and the Gospel is preached to the poor* (Mat. 11:5)

It seems that GMat is using *być* for both aspects of participles, and occasionally the frequentative *bywać* with iterative meaning (in example (18) the Gospel is preached to the poor again and again, not just once), whereas WMat uses either *być* or *zostać* with perfectives and *być* with imperfectives, as is the case in Modern Polish. Splitting the concordance into matches with perfective and imperfective participles clearly shows the limitation of *zostać* to the perfective aspect: the imperfective concordance in Table 5.4 has only *być*, but the perfective one in Table 5.5 shows both, as in Table 5.3:

GMat	WMat	A	B	C	MI3
<i>być</i>	<i>być</i>	13	15	7	6.137

Tab. 5.4: Single correspondence next to imperfective passive participles.

GMat	WMat	A	B	C	MI3
<i>być</i>	<i>być</i>	169	143	71	12.709
<i>być</i>	<i>zostać</i>	169	29	18	9.071

Tab. 5.5: Two correspondences next to perfective passive participles.

In fact, all 18 cooccurrences with *zostać* (column C) are accounted for in the perfective concordance of Table 5.5, and the amount of parallel imperfective predicative passive participles is generally very low, with only 7 cooccurrences appearing next to *być*. The fact that the *być* : *być* correspondence is still more frequent in the perfective aspect than *być* : *zostać* is surprising and anomalous for standard modern Polish, which uses *zostać* with perfectives much more often (cf. Swan, 2002: 312-314). This deviation may perhaps be attributable to the relative conservatism of the newer translation, which could have been influenced by the source language(s), or by the language of the Gdansk Bible itself. Nonetheless, the typical modern use of *zostać* is clearly in evidence as well.

<sup>23</sup> The form *jest* 'is' is the suppletive present tense form of the same lemma *być* 'to be' discussed above.



## 5.5 Summary and Evaluation

In this chapter I have examined grammatical categories in isolation and in the context of adjacent tokens. It has been possible to detect the decline in the use of the active past participle in favor of finite past tense verbs ('having done' > 'did'), the overall tendency to use genitives of proper nouns instead of congruent possessive adjectives ('Davidian' > 'David's') and the use of the copula verb *zostać* 'to become' with predicative perfective participles, which does not occur in the old text. In every case the corpora have also shown identical, unchanged forms for each of the examined constructions (in the latter case even more often than not), but the time of WMat's authorship has inevitably left its Modern Polish marks on the text, despite its conservative nature.

Methodologically, the results also show that non-lexical annotation information on the token level, such as part-of-speech, and on the syntagm level, such as congruence, can be used to discriminate target token populations in order to produce verse aligned concordances and find the most statistically significant correspondences between the corpora. A very positive aspect of this process is its relative open-endedness. Although the results given by a query are always dependent on the way it is formulated, the queries phrased in this chapter imposed relatively few constraints on the parallels they might have revealed. For example, the most significant correspondence of the label signifying an active past participle happens to be a finite verb, but it could have been any other label as far as the query's formulation is concerned. The methods employed here are exploratory: the correspondence table between all possible stripped items can be examined, and significant matches can be studied further by refining queries to address different, possibly discriminating factors (e.g. adjacency to a passive participle in section 5.4, and on an even finer level, the category of aspect within such adjacent bigrams).

By examining runners-up, it is also possible to find the items that are in the closest competition in WMat for a grammatical slot as defined by the occurrence of a single category in parallel in GMat (as we have seen, usually between retention of the same, presumably more archaic construction found in GMat vs. a more modern alternative). This may not reveal all the different competitions which are in action in the language of WMat, since they may not correspond to categorical choices in GMat, but on the other hand, it allows a direct examination in parallel of what has happened to

constructions that do occur in the language of GMat (recall the difference between the frequency graphs in section 5.3 and the paired query results).

While it is again difficult to quantitatively evaluate results in the absence of a formal gold standard of correspondences one expects to find, it is evident that the results presented in this chapter all correspond to actual historical processes in Polish as documented in traditional historical grammars. Whether or not these results are distributed in a representative way is impossible to answer without a contemporary Middle Polish reference corpus, though it is safe to say that the view of the Bible as a conservative text (cf. section 1.2) is probably justified in light of previous treatments of these phenomena, and one can expect that different texts might have shown much more progressive statistics in favor of the more modern alternatives described in this chapter. On the other hand, the fact that the more archaic constructions are often overridden in WMat is all the stronger evidence that they are going out of use, with the Bible representing one of their last bastions. It can also be expected that less frequent phenomena, as well as less clear cut correspondences, will not be demonstrable in the small corpus used in this study: especially syntactic phenomena notoriously often require a very large corpus to substantiate.

Finally, with regard to the limitations of the tools used in this chapter, it is clear that the querying possibilities of a flat, unchunked corpus are very limiting for the study of complex syntagms. Just as non copula-adjacent passive predications had to be given up in section 5.4, so are longer  $n$ -grams and of course discontinuous, or variable length constructions very difficult or impossible to capture with these techniques. Also, the SQL infrastructure, which perhaps surprisingly suffices for all the studies conducted in this work, is at its very limit when syntagms and their internal relations come into play. Some thoughts on discontinuous and hierarchical data within the parallel concordance methodology presented so far will be given in chapter 6 for future study.

## 6 Conclusion and Outlook

In this study I have explored the uses of a parallel diachronic corpus for the extraction of changes forming part of the historical grammar of the Polish language, or more narrowly, of the Polish language as it is expressed within the genre or sublanguage of biblical scripture, in the last four centuries. The examined areas have encompassed some of the central points of historical change, including changes in inflectional morphology, word formation and vocabulary, as well as the use of different grammatical categories on a single- and multi-token level. While only certain facets of each area could be studied in the scope of this work, it is highly likely that the same techniques used here could be extended to cover changes in adjectival, pronominal or verbal morphology, as well as nominal lexis and a variety of other phenomena involving grammatical categories and changes in syntax (the latter subject to some limitations which will be taken up below).

The results of the investigations that were carried out show that although biblical language is rather conservative quantitatively, especially in so far as the constancy of lexis can be judged from the high proportion of identical or partly identical lemma pairs found in chapter 4, most changes described in traditional historical grammars can probably be found, at least qualitatively, by comparing the two diachronically disparate Bible texts automatically. In other words, while there is always a substantial overlap component in each area (forms with the same suffixes, lemmas, and constructions), differences between the language stages are bound to crop up – otherwise the texts could not be said to belong to these language stages – and these can be found using the techniques outlined in this work. The added value of using these methods goes beyond the facilitation of a manual search for phenomena which are already known, but extends to a reproducible, data-based, and most importantly quantitative account of these phenomena, descriptions of which may otherwise often be restricted to qualitative statements that may be subjective to a greater extent.

Absent from this study is an account of the historical phonology of Middle and Modern Polish, which was made impossible by the normalized orthography used in the edition of the older text that has been used here. It is not implausible that this area could also be studied successfully using automatic parallel corpus-based techniques, in

particular taking advantage of the fruits of existing research on automatic cognate identification (see e.g. Bergsma and Kondrak, 2007). Given a function determining cognate status to filter out non-cognates from a parallel concordance of non-normalized pairs, it is conceivable that micro-correlations on a character- (or ideally phoneme-)based level could be used to extract sound change correspondences, and even their conditioning environments.

The methods used in this work can be divided into two major types: the first, exploiting the extremely comparable contents of the parallel corpus, examines the unpaired binary occurrence or non-occurrence of identical and minimally different tokens between the corpora, and their discrete quantifiable distributions, as illustrated by the study of inflectional morphology in chapter 3. This type of approach is especially suitable for very well attested phenomena, where minor stylistically or otherwise determined but non-systematic differences in distribution are not expected to adversely skew results, and it can be reasonably assumed that almost any change that has taken place will be in evidence in the data. The lack of differences between the corpora with respect to subject matter, register or genre, thanks to the parallel text, ensures that observable effects in the distributional relationships of categories are mainly due to diachronic factors. The completeness of the coverage of morphological change using this technique is therefore owing to the fact that the categories are very frequent, and the highly comparable corpora are conducive to instances of minimal token pairs with only the changed properties mismatching.

The second type of method harnesses the parallel alignment of the corpora to extract consistent correspondences between particular items. This technique, which can operate unsupervised on any annotation layer or layers, outputs a ranked concordance of the correspondences for each item, which can be refined to target specific items or types of items by applying annotation criteria to filter them, as well as criteria computed from the concordance itself, such as correlation measures (MI3 or similar measures) and string manipulation functions on either each one of the items (as in SoundEx-like phonetic reduction functions) or both (as in Levenshtein distance). In essence, any categorical distinction whose occurrence or non-occurrence can be operationalized to produce a concordance of the items showing this category and its respective values, alongside the

parallel identifiers of the segments where they occur, can be correlated automatically across corpora. The question is only how to phrase the queries to retrieve exactly the items we are interested in. The truism thus holds that the corpus can and will yield more or less exactly what is annotated in it, but the parallel extraction will allow this to be done quickly, effortlessly, and possibly using much less data, as the small size of this corpus suggests.

Especially this last point is important for the feasibility of using parallel corpora for historical linguistics: it is precisely in historical linguistics that one is interested in making the most of smaller corpora, and although Bible data is far from optimal, this case study has hopefully shown that even this text inevitably contains diachronically significant variation, as attested by the corpus-based results matching traditional descriptions, which are based on the entire historical corpus of Polish. However the added value of the corpus driven approach is twofold: it allows a less biased, more open-ended mode of research, where the researcher truly does not know what results he or she will find, and it delivers quantifiable data of the sort which is impossible to gather otherwise. And while the relevance of a linguistic statement on, for example, the amount of retained or altered verbal lemmas between two diachronically disparate texts is debatable with regard to the state of affairs in the so-called ‘general language’, it is nevertheless operationalizable, and as such of interest at least as a case study. For instance, the relative readability of the Gdansk Bible for a Modern Polish reader could be assessed and accounted for in this way, or compared across different parts of the text (e.g. New vs. Old Testament) or with other texts.

As for future work within this paradigm, there are many open questions and some exciting prospects. The most burning issue for the data already gathered is its relationship with the ‘general language’. Despite the great potential of parallel corpora, their comparatively small size, and in the diachronic case often their exclusively religious language (though the latter problem extends beyond parallel corpora in historical linguistics), limit them to a small sublanguage. At the same time, it is very clear that even some of the variation identified between the corpora in this work is entirely due to stylistic or even completely coincidental variation, e.g. the selection of freely alternating default prefixes in *spytać* vs. *zapytać* ‘to ask’. On the other hand, other differences are

very clearly diachronically motivated, as in the case of more or less archaic lemmas, which have in many cases completely gone out of use, or even more so in the categorical, sweeping replacement of an inflectional suffix. The limited parallel corpus is incapable of distinguishing the various factors responsible for distinct types of differences such as stylistic variation, diatopic influences and, quite possibly, the peculiarities of translated text. An empirical evaluation and further categorization of these cases can only be achieved using more corpus data. On the one hand, larger, heterogeneous, and non-parallel (but ideally comparable) monolingual diachronic corpora should be used in order to determine the changes in the overall frequencies of the items in question, and on the other hand, multiple contemporary versions of the same parallel texts could be used to filter out stylistic and other non-diachronic kinds of variation. In the case of the Bible it is especially feasible to compare multiple translations, thereby also creating a better temporal resolution, and gaining the ability to dismiss differences that contrast with fewer texts or only one, and accept those differences that are common to multiple contemporaneous corpora as truly attributable to the diachronic dimension. This is a difficult and resource consuming task, but in historical linguistics it has no alternatives, since non corpus-based approaches cannot be assigned the validity that they have for living languages. At the same time, the emergence of more complete national language historical corpora makes it not unimaginable that the sum total of e.g. old Bible translations will become available in an electronic, annotated format for many languages. These, in conjunction with the techniques explored here and others, could be used to create data-based, fine grained historical lexica, supplying for example on-line dictionaries with quantifiable distributional and correspondence information which could then be visualized to illustrate and compare the development of different items.

The other main point requiring attention is the further development of parallel corpus-based techniques for historical purposes. Here there is much to be said for employing more recent techniques developed for machine translation and parallel terminology extraction, such as example-based machine translation (EBMT, see Somers, 1999 and to appear, and Cicekli and Güvenir, 2001 among many others). This direction of techniques allows the automatic acquisition of ‘translation templates’ with variable components from a parallel corpus, creating a concordance of possibly discontinuous

constructions with empty slots, which would be highly valuable for the study of syntactic correspondences. The study of discontinuous and hierarchical constructions is in general a major challenge for the approach laid out here, as it is also, not surprisingly, for machine translation and related fields. Some easier headway can likely be achieved using chunking and a subsequent parallel alignment of chunks representing phrase structure. Patterns of chunks can then be matched to investigate more complex constructions, such as argument structures. A further level of sophistication could be reached by correlating explicitly tagged nodes in a syntactically parsed corpus, though here the syntactic knowledge required to parse, either automatically or manually, already adds a substantial further dimension of theory-dependent interpretation which may possibly reduce the open-endedness of this approach. The scarce availability of parsed historical corpora is also an obstacle to such endeavors. On the other hand, it might be possible to address both these difficulties by inducing parallel syntactic structures from a raw parallel corpus, or else using a partially annotated seed corpus, by applying a bootstrapping approach (see Kuhn, 2005).

A final point of improvement already mentioned in chapter 4 would be the use of semantic sense annotation, on either a manual or context-based (semi-)automatic basis, which would allow the distinction of multiple developments for the same orthographic item. Using further parallel texts in more distant languages could also be used to distinguish senses for this purpose (cf. Resnik and Yarowsky, 2000 and Dyvik, 2002), which in the case of the Bible text is especially feasible. Sense annotation might even be helpful already for the study of inflectional morphology: semantic tagging could distinguish such expressions as *wyjsć za mąż* ‘marry (a man)’, with the fossilized ancient accusative of the word for ‘husband’ (cf. section 3.2), from other uses of this word on the basis of either monolingual context or different translations, though arguably this sense difference should perhaps be resolved already at the level of tokenization, where the phrase could form a multi-word unit.

In conclusion, I hope to have shown that many elements of change in historical grammar and lexis can be identified automatically or semi-automatically using a parallel diachronic corpus, which should ideally be at least lemmatized, though grammatical and morphological suffix annotation have both proven to be valuable here. The techniques



which are needed in order to extract the relevant correspondences between texts in different language stages are already known from computational linguistics and can be adapted for this purpose with relative ease. Their implementation too has been shown to be achievable, for a flat corpus, using little more than SQL and some basic string comparison and editing functions to interpret and classify results. The power behind these techniques lies in the excellent comparability and alignment of the parallel text, which makes it possible to illuminate differences using minimal pairs and to know not just how frequent a phenomenon is or was in each language stage, but also more precisely what is replacing what and to what extent.

## 7 Bibliography

Baroni, M. and Bernardini, S. (2006), “A New Approach to the Study of Translationese: Machine-Learning the Difference between Original and Translated Text”. *Literary and Linguistic Computing* 21(3), 259-274.

Beekes, R. S. P. (1995), *Comparative Indo-European Linguistics: An Introduction*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Bergsma, S. and Kondrak, G. (2007), “Alignment-Based Discriminative String Similarity”. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, 656–663.

Berry-Rogghe, G. L. M. (1973), “The Computation of Collocations and their Relevance in Lexical Studies”. In: Aitken, A. J., Bailey, R. and Hamilton-Smith, N. (eds.), *The Computer and Literary Studies*. Edinburgh: Edinburgh University Press, 103-112.

Biber, D. (1993), “Representativeness in Corpus Design”. *Literary and Linguistic Computing* 8(4), 243-257.

Bielfeldt, H. H. (1961), *Altslawische Grammatik, Einführung in die slawischen Sprachen*. Halle: Veb Max Niemeyer Verlag.

Cicekli, I. and Güvenir, H. A. (2001), “Learning Translation Templates from Bilingual Translation Examples”. *Applied Intelligence* 15, 57-76.

Curzan, A. (to appear), “Historical Corpus Linguistics and Evidence of Language Change”. In: Lüdeling, A. and Kytö, M. (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Cysouw, M. and Wälchli, B. (to appear), “Parallel Texts: Using Translational Equivalents in Linguistic Typology”. Special issue of *STUF*.

Daille, B. (1995), *Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering*. Unit for Computer Research on the English Language Technical Papers 5, Lancaster University.

Długosz-Kurczabowa, K. and Dubisz, S. (2006), *Gramatyka historyczna języka polskiego*. Warsaw: Wydawnictwa Uniwersytetu Warszawskiego.

Dunning, T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence". *Computational Linguistics* 19(1), 61-74.

Dyvik, H. (2002), "Translations as Semantic Mirrors. From Parallel Corpus to WordNet". In: Aijmer, K. and Altenberg, B. (eds.), *Advances in Corpus Linguistics. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23) Göteborg 22-26 May 2002*. Amsterdam: Rodopi, 311-326.

Ehrman, B. D. (2005), *Misquoting Jesus. The Story Behind Who Changed the Bible and Why*. New York: HarperSanFrancisco.

Evert, S. (2005), *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, University of Stuttgart.

Fung, P. (1998), "A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora". *Lecture Notes in Artificial Intelligence* 1529, 1-17.

Hansen-Schirra, S. and Teich, E. (to appear), "Corpora in Human Translation". In: Lüdeling, A. and Kytö, M. (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Jespersen, O. (1924), *The Philosophy of Grammar*. London: George Allen & Unwin.

Klemensiewicz, Z. (1999), *Historia języka polskiego*. Wydanie siódme, uzupełnione. Warsaw: Wydawnictwo Naukowe PWN.

Klemensiewicz, Z., Lehr-Spławiński T. and Urbańczyk, S. (1955), *Gramatyka historyczna języka polskiego*. Warszawa: Państwowe Wydawnictwo Naukowe.

Koziara, S. (2001), *Frazeologia biblijna w języku polskim*. Kraków: Wydawnictwo Naukowe Akademii Pedagogicznej.

Kuhn, J. (2005), "An Architecture for Parallel Corpus-Based Grammar Learning". In: Fisseni, B., Schmitz, H-C., Schröder, B. and Wagner, P. (eds.), *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*. Frankfurt am Main: Peter Lang, 132-144.

Labov, W. (1994), *Principles of Linguistic Change. Volume 1: Internal Factors*. Oxford, UK and Cambridge, MA: Blackwell.

Levenshtein, V. I. (1966), "Binary Codes Capable of Correcting Deletions, Insertions, and Reversals". *Soviet Physics Doklady* 10, 707–710.

Lüdeling, A. (2007), "Das Zusammenspiel von qualitativen und quantitativen Methoden in der Korpuslinguistik". In: Zifonun, G. and Kallmeyer, W. (eds.), *Jahrbuch des Instituts für deutsche Sprache 2006*. Berlin: de Gruyter, 28-48.

Mańczak, W. (1956), "Ile jest rodzajów w polskim?". *Język Polski*, XXXVI(2), 116-121.

Martinet, A. (1962), *A Functional View of Language*. Oxford: Clarendon Press.

Nurmi, A. (2002), "Does Size Matter? The Corpus of Early English Correspondence and its Sampler". In: Raumolin-Brunberg, H., Nevala, M., Nurmi, A. and Rissanen, M. (eds.), *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen*. Mémoires de la Société Néophilologique de Helsinki LXI. Helsinki: Société Néophilologique, 173-184.

Oakes, M. P. (1998), *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Pearce, D. (2002), "A Comparative Evaluation of Collocation Extraction Techniques". In: *Third International Conference on Language Resources and Evaluation, May, 2002*. Las Palmas, Canary Islands, Spain.

Pisarkowa, K. (1984), *Historia składni języka polskiego*. Wrocław et al.: Polska Akademia Nauk.

Proctor, P. (ed.) (1978), *Longman Dictionary of Contemporary English*. Harlow and London: Longman.

Przepiórkowski, A. (2003), "A Hierarchy of Polish Genders". In: Bański, P. and Przepiórkowski, A. (eds.), *Generative Linguistics in Poland: Morphosyntactic Investigations*. Warsaw: Institute of Computer Science, Polish Academy of Sciences, 109-122.

Przepiórkowski, A. (2004) *The IPI PAN Corpus, Preliminary Version*. Warsaw: Institute of Computer Science, Polish Academy of Sciences.

Przepiórkowski, A. and Woliński, M. (2003), "A Flexemic Tagset for Polish". In: *Proceedings of the Workshop on Morphological Processing of Slavic Languages, EACL 2003*. Budapest. Available from: <http://nlp.ipipan.waw.pl/~adamp/Papers/2003-eacl-ws12/ws12.pdf>.

Resnik, P. and Yarowsky, D. (2000), "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation". *Natural Language Engineering* 5(2), 113-133.

Resnik, P., Broman Olsen, M. and Diab, M. (1999), "The Bible as a Parallel Corpus: Annotating the 'Book of 2000 Tongues'". *Computers and the Humanities* 33, 129-153.

Rissanen, M. (1989), "Three Problems Connected with the Use of Diachronic Corpora". *ICAME Journal* 13, 16-19.

Rissanen, M. (to appear), "Corpus Linguistics and Historical Linguistics". In: Lüdeling, A. and Kytö, M. (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Rospond, S. (2003), *Gramatyka historyczna języka polskiego, z ćwiczeniami*. Warsaw: Wydawnictwo Naukowe PWN.

Siatkowski, J. (1970), *Bohemizmy fonetyczne w języku polskim*. Vol. 2. Wrocław: Zakład Narodowy im. Ossolińskich.

Simard, M., Foster, G., Hannan, M-L., Macklovitch, E. and Plamondon, P. (2000), "Bilingual Text Alignment: Where do we Draw the Line?". In: Botley, S. P., McEnery, A. M. and Wilson, A. (eds.), *Multilingual Corpora in Teaching and Research*. Amsterdam / Atlanta, GA: Rodopi.

Śmiech, W. (1986), *Derywacja prefiksalna czasowników polskich*. Prace wydziału I – językoznawstwa, nauki o literaturze i filozofii 87. Wrocław et al.: Łódzkie Towarzystwo Naukowe.

Somers, H. (1999), "Review Article: Example-based Machine Translation". *Machine Translation* 14, 113-157.

Somers, H. (to appear), "Corpora and Machine Translation". In: Lüdeling, A. and Kytö, M. (eds.), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter.

Swan, O. E. (2002), *A Grammar of Contemporary Polish*. Bloomington, IN: Slavica Publishers.

Swan, O. E. (no date), *A Learner's Polish-English Dictionary*. First Preliminary Edition. CD and Web Version. Available at <http://polish.slavic.pitt.edu/dictionary.pdf>.

Taube, M. (1980), "On the Penetration of the Perfect into the Russian Narrative System". *Russian Linguistics* 5, 121-131.

Vaillant, A. (1950-1977), *Grammaire comparée des langues slaves*. Paris: Éditions Klincksieck.

Wackernagel, J. (1882), „Über ein Gesetz der indogermanischen Wortstellung“. *Indogermanische Forschungen* 1, 333-436.

Wiśniewska, H. (1994), *Kulturalna polszczyzna XVII wieku (na przykładzie Zamościa)*. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.

Włodarczyk, A. and Włodarczyk, H. (2001), “La préfixation verbale en polonais. I. Le statut grammatical des préfixes, II. L’Aspect perfectif comme hyper-catégorie”. *Études cognitives / Studia kognitywne* 4, 93-120.

Wolf, H. (1996), “Einführung”. In: Wolf, H. (ed.), *Luthers Deutsch. Sprachliche Leistung und Wirkung*. Frankfurt am Main: Peter Lang, 9-29.

Woliński, M. and Przepiórkowski, A. (2001), *Projekt anotacji morfosyntaktycznej korpusu języka polskiego*. IPI PAN Reports 938, December 2001. Available from: <http://nlp.ipipan.waw.pl/~adamp/Papers/2001-tagset/ipi938.pdf>.

Yarowsky, D. and Ngai, G. (2001), “Inducing Multilingual POS Taggers and NP Brackets via Robust Projection across Aligned Corpora”. In: *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics*. Pittsburgh, PA, 200-207.

Zeldes, A. (2006) Abstracting Suffixes: A Morphophonemic Approach to Polish Morphological Analysis. In: *Proceedings of Konvens'06, Konstanz, 4-7 October, 2006*. Konstanz, 151-158. An extended version of this article is set to appear in a special issue of *Zeitschrift für Sprachwissenschaft*.



## Appendix A – List of Abbreviations

The following abbreviations are used especially in the glosses of examples, and occasionally elsewhere. For convenience and to save space, in some cases a gloss gives a literal translation instead of a categorical coding, where the latter is not relevant, e.g. *sent* (and not necessarily *send-PAST*), or else both a literal translation and coding, e.g. *married-PASS-PFV* (and not *marry-PASS-PFV*). For more on part-of-speech tags see section 2.4.

1 – first person

2 – second person

3 – third person

ACC – accusative case

AUX – auxiliary verb

DAT – dative case

EMPH – emphatic particle

F – feminine gender

FREQ – frequentative verb

GEN – genitive case

IMPFV – imperfective

INST – instrumental case

LOC – locative case

M – masculine gender

MA – masculine animate gender

MI – masculine impersonal gender

MP – masculine personal gender

N – neuter gender

NOM – nominative case

nV – non virile plural (agreeing with a plural without masculine personal gender)

PARTPF – perfect participle

PASS – passive

PAST – past tense

PFV – perfective

PL – plural

POS – part-of-speech

PPA – active past participle

PREP – preposition

PRES – present tense

REFL – reflexive pronoun

SG – singular

V – virile plural (agreeing with plural containing a masculine personal gender)

VFIN – finite verb (including past tense, originally participle based forms)

VOC – vocative case

VT – transitive verb

## Appendix B – the *agr* Function

The following function, which is based on MSSQL string functions, receives two annotated tokens GA and GB as arguments and returns either the string "agr" if the tokens are congruent, or the empty string if they are not.

```
iif
(
    (
        (left(GA.pos,1) = "S") and
        (
            ((instr(GB.pos,"Adj")>0) or (instr(GB.pos,"Pro")>0)
            or (instr(GB.pos,"Part")>0) and not (GA.pos =
            "VPartPastAct"))
            and
            GA.case=GB.case
            and
            (
                GA.gend=GB.gend
                or
                (
                    (GA.gend="MP" and GB.gend="M") and
                    (GA.num=GB.num)
                )
                or
                (
                    GA.num = "pl" and GB.num="pl" and
                    (
                        (GA.gend="MP" and GB.gend="V") or
                        (
                            (GA.gend="MI" or GA.gend="F"
                            or GA.gend="N" or
                            GA.gend="MA")
                            and
                            (GB.gend="nV")
                        )
                    )
                )
            )
        )
    )
    or
    (
        (left(GB.pos,1) = "S") and
        (
            ((instr(GA.pos,"Adj")>0) or (instr(GA.pos,"Pro")>0)
            or (instr(GA.pos,"Part")>0) and not (GA.pos =
            "VPartPastAct"))
            and
            GB.case=GA.case
            and
            (

```

```

GB.gend=GA.gend
or
(
    (GB.gend="MP"    and    GA.gend="M")    and
    (GB.num=GA.num)
)
or
(
    GB.num = "pl" and GA.num="pl" and
    (
        (GB.gend="MP" and GA.gend="V") or
        (
            (GB.gend="MI" or GB.gend="F"
            or GB.gend="N" or
            GB.gend="MA")
            and
            (GA.gend="nV")
        )
    )
)
)
)
)
or
(
    (
        (GA.pos = "VFin" and GA.pers=3 and GA.num=GB.num and
        GB.case="nom")
        and
        (
            (GA.gend=GB.gend)
            or
            (GA.gend="M" and left(GB.gend,1)="M")
            or
            (
                (GA.num="pl" and GB.num="pl")
                and
                (
                    (
                        GA.gend="nV"
                        and
                        (
                            GB.gend="F" or
                            GB.gend="N" or
                            GB.gend="MI" or
                            GB.gend="MA"
                        )
                    )
                    )
                )
            )
        )
    )
)
)
)

```

```

        )
        or
        (isnull(GA.gend))
    )
)
or
(
    (GB.pos = "VFin" and GB.pers=3 and GB.num=GA.num and
    GA.case="nom")
    and
    (
        (GB.gend=GA.gend)
        or
        (GB.gend="M" and left(GA.gend,1)="M")
        or
        (
            (GB.num="pl" and GA.num="pl")
            and
            (
                (
                    GB.gend="nV"
                    and
                    (
                        GA.gend="F" or
                        GA.gend="N" or
                        GA.gend="MI" or
                        GA.gend="MA"
                    )
                )
                or
                (
                    GB.gend="v"
                    and
                    (
                        GA.gend="MP"
                    )
                )
            )
        )
    )
    or
    (isnull(GB.gend))
)
)
)
, "agr", ""
)

```